**Evaluating the Effectiveness of SMOTE and SMOTE-ENC in Class Imbalance Medical Data**

**Literature Review**

**Oversampling Techniques for COPD Diagnosis and Predictive Analytics**

Machine learning diagnostic effectiveness declines because COPD subtype classification faces difficulties due to uneven class distribution. Data balancing stands as an important factor in developing dependable systems which deliver unbiased predictions. Oversampling methods together with SMOTE-ENC stand as crucial solutions for handling imbalanced datasets that contain mixed features like those in the ExaSens dataset. With the correct combination of preprocessing steps and oversampling strategies machine learning models achieve better predictive performance and enable early disease detection while creating customized treatment plans (Mukherjee and Khushi, 2021). The ability of these systems to manage multiple types of datasets reveals their significant capability to transform COPD classification alongside demonstrating how data-driven techniques enhance healthcare results.

**Synthetic Oversampling Techniques for Mixed-Feature Datasets**

The Synthetic Minority Oversampling Technique (SMOTE) combined with other synthetic oversampling methods remains the standard approach to create synthetic minority class samples for class imbalance correction. Standard SMOTE works exclusively with continuous data which restricts its application on datasets that include both nominal and continuous features. SMOTE-ENC builds upon SMOTE by encoding nominal features together with continuous ones to achieve effective oversampling for mixed-feature datasets (Chowdhury, Ayon and Hossain 2024). Medical research benefits greatly from this technique because datasets in the field often combine categorical information like smoking status with continuous variables such as biomarkers. SMOTE-ENC maintains mixed data structure integrity while improving machine learning model outcomes on difficult datasets.

**Literature Review of Class Imbalance in Dataset**

The study (Cu *et al.,* 2024) suggested Improved SMOTE (I\_SMOTE) and Improved ADASYN (I\_ADASYN) as new methods for data imbalance resolution based on revised versions of existing SMOTE and ADASYN algorithms. A novel candidate selection method that combined similarity measurement with a roulette wheel selection mechanism created synthetic data samples through these approaches. Evaluation for educational datasets occurred using public datasets alongside data collected from a university in Vietnam. The study showed I\_SMOTE and I\_ADASYN methods achieved better results by enhancing recall and precision levels along with F1-score for minority classes. The improved versions of SMOTE and ADASYN boosted ROC area performance by 6.6% and 8.0% above the original methods.

The study (Firat Atay *et al.,* 2024) developed a relational classification model for DTC recurrence prediction that combined clinical information with pathological and follow-up data. The research analyzed a balanced 550 sample dataset from 15 years of DTC data which included 13 clinicopathological variables. The study managed the recurrence status class imbalance through the application of SMOTE-NC. Researchers implemented a hybrid methodology that combined classification algorithms and association rule mining using RCAR and CBAR methods. Independent predictors of recurrence emerged through binomial logistic regression analysis. Different performance measurements like accuracy together with sensitivity and F1 score served as evaluation tools. RCAR achieved 96.7% accuracy which surpassed CBAR performance while key predictors of recurrence were papillary pathology and incomplete response according to association rules.

A table with numbers and text

Description automatically generated

**Figure Performance results of classifiers (Firat Atay *et al.,* 2024).**

The study (Chowdhury, Ayon and Hossain, 2024) explored several data augmentation strategies such as ENN, SMOTE-N, SMOTE-ENN and SMOTE-Tomek to solve issues caused by imbalanced datasets. Four machine learning algorithms, logistic regression, AdaBoost, gradient boosting and random forest were tested for effectiveness on various sampling methods. Analysis utilized the BRFSS dataset where individuals with diabetes formed a minority class amid substantial class imbalance. ENN combined with gradient boosting demonstrated the highest recall at a 14.2% improvement level while AdaBoost and logistic regression with ENN produced recall enhancements of 13.3% and 13.1% respectively. With ENN soft voting classifiers reached 71.4% recall and 78.9% AUC which demonstrated strong capabilities for detecting minority class elements.

This study (Hamida *et al.,* 2024) introduced a new hybrid diagnostic system which combines random forest and deep neural network algorithms to tackle skin disease detection problems. The study implemented data augmentation alongside balancing methods to enhance both model performance and its ability to generalize. Researchers trained and evaluated their model using the HAM10000 dataset which contains diverse dermatoscopic images. The hybrid model first applied the RF algorithm for preliminary patient diagnosis based on reported symptoms and then used DNN analysis of lesion images to produce exact diagnoses. The system underwent hyper-parameter optimization to achieve optimal performance tuning. The hybrid approach achieved top performance with a classification accuracy rate of 96.8%.

A table with numbers and text

Description automatically generated

**Figure Result analysis (Hamida *et al.,* 2024)**

Through Synthetic Minority Oversampling (SMOTE) data balancing and ensemble stacking learning techniques this study (Rofik *et al.,* 2024) improved credit assessment accuracy. The model design incorporated Random Forest, SVM, Extra-Tree Classifier as initial learning algorithms together with XGBoost serving as the higher-level corrective learning mechanism. Data collection proceeded through preprocessing steps before oversampling occurred followed by modeling and evaluation. For testing purposes cross-validation was applied to the German Credit dataset. The developed model proved to be highly effective producing performance results including 83.21% accuracy along with 79.29% precision while reaching 91.78% recall and achieving an F1-score of 85.08%. The research proved that SMOTE along with stacking ensemble learning methods can improve credit scoring task performance outcomes.

The study (Chachoui *et al.,* 2024) suggested a unique solution called Equi-Fused-Data-based SMOTE to tackle imbalanced educational datasets in machine learning and improve learning outcomes. The study's dataset included data from 2,176 first-year novice programming students, that underwent preprocessing and oversampling with SMOTE alongside SMOTE Borderline, SMOTE-ENN and ADASYN. Through data diversity maximization from merged oversampled datasets researchers evaluated their results using a balanced bagging model. The method displayed high effectiveness by producing 93.85% accuracy, 92.86% precision and recall along with an F1-score and 98.08% AUC which proves its strong performance.

A screenshot of a computer

Description automatically generated

**Figure Performance on SMOTE balanced dataset (Chachoui *et al.,* 2024).**

Cyclists' subjective evaluations were assessed by the study (Zeng et al., 2024) through a novel SSB framework of Public Security, Traffic Safety, Scenic Beauty which utilized street view imagery (SVI) supported by volunteer rating input. The study used both Kmeans Synthetic Minority Over-Sampling Technique (Kmeans-SMOTE) and Random Forest (RF) classifier to manage volunteer ratings imbalance and better distinguish positive from negative perceptions. The Kmeans-SMOTE-RF model showed substantial performance gains in Area Under the Curve (AUC) by achieving 0.327 improvement for public safety and enhancements of 0.2 for traffic safety and 0.209 for scenic beauty when compared to the Random Forest model alone.

While oversampling techniques help correct dataset imbalances, they tend to generate noise and ambiguous class boundaries which cause models to overfit the training data. The study (Zhang, Deng and Wei, 2024) developed the FSDR-SMOTE methodology to address these issues by merging Random-SMOTE with Feature Standard Deviation analysis. The method begins by eliminating noisy data points through Tukey's criterion before performing feature standard deviation analysis to locate samples which it then marks as either boundary or safety samples. K-means clustering splits minority class samples into multiple sub-groups to produce new samples which include characteristics of random samples boundary samples and cluster centers. Research data from twenty UCI benchmark datasets revealed that FSDR-SMOTE achieved an average F-measure result of 93.31% together with G-mean and MCC scores of 93.16% and 86.53% respectively which proves its effectiveness.

**Table Summary of Literature Studies**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Study** | **Dataset** | **Aim** | **Data Balancing Technique** | **Key Findings/Results** |
| Cu *et al.,* (2024) | Educational datasets, Vietnam | Improve model performance for minority classes in educational datasets | I\_SMOTE, I\_ADASYN | I\_SMOTE and I\_ADASYN improved recall, precision, F1-score for minority classes. ROC area improved by 6.6% and 8.0%, respectively. |
| Firat Atay *et al.,* (2024) | 550 DTC samples (15 years) | Predict recurrence of DTC using clinicopathological data | SMOTE-NC | RCAR achieved 96.7% accuracy, sensitivity, and F1-score. Papillary pathology and incomplete response were key predictors. |
| Chowdhury, Ayon and Hossain (2024) | BRFSS dataset (Diabetes) | Improve performance in detecting diabetes in imbalanced data | ENN, SMOTE-N, SMOTE-ENN, SMOTE-Tomek | ENN with gradient boosting improved recall by 14.2%. Soft voting classifiers with ENN reached 71.4% recall, 78.9% AUC. |
| Hamida *et al.,* (2024) | HAM10000 dermatoscopic images | Improve skin disease detection using hybrid machine learning models | Data augmentation, balancing methods | Hybrid model achieved 96.8% classification accuracy. |
| Rofik *et al.,* (2024) | German Credit dataset | Improve credit scoring prediction accuracy | SMOTE, ensemble stacking | Accuracy: 83.21%, Precision: 79.29%, Recall: 91.78%, F1-score: 85.08%. SMOTE + stacking improved credit scoring performance. |
| Chachoui *et al.,* (2024) | 2,176 students (novice programmers) | Improve learning outcomes by balancing data for educational datasets | SMOTE, SMOTE Borderline, SMOTE-ENN, ADASYN | Accuracy: 93.85%, Precision: 92.86%, Recall: 92.86%, AUC: 98.08%. High effectiveness in improving learning outcomes. |
| Zeng *et al.,* (2024) | Street view imagery (SVI) | Enhance cyclists' perception assessment for public safety, traffic, and beauty | Kmeans-SMOTE | AUC improved by 0.327 for public safety, 0.2 for traffic safety, 0.209 for scenic beauty compared to RF model. |
| Zhang, Deng and Wei (2024) | 20 UCI benchmark datasets | Improve classification accuracy for imbalanced datasets | FSDR-SMOTE | FSDR-SMOTE achieved F-measure: 93.31%, G-mean: 93.16%, MCC: 86.53%. Effective for imbalanced datasets. |

**Comparative Analysis**

Previous studies reveal diverse approaches for handling data imbalance issues in the machine learning domain by applying them specifically to medical and educational contexts. Cu *et al.,* (2024) developed I\_SMOTE and I\_ADASYN methods which demonstrated superior performance than traditional SMOTE and ADASYN through enhanced recall, precision and F1-score results within educational datasets. First Atay *et al.,* (2024) concentrated on prediction imbalances for DTC recurrence by using SMOTE-NC to achieve top results in both accuracy and recall. Rofik *et al.,* (2024) applied SMOTE with ensemble stacking to credit assessment which resulted in an 83.21% accuracy rate. According to research conducted by Hamida *et al.,* (2024) and Chowdhury *et al.,* (2024) SMOTE techniques delivered significant advancements in detecting diseases and classifying diabetes. Zhang, Deng, and Wei (2024) developed FSDR-SMOTE to address the issue of overfitting and fuzzy boundaries in oversampling but their method does not work with medical datasets containing both nominal and continuous features. Existing methods which boost recall through oversampling lack specific accommodation for mixed feature datasets containing both nominal and continuous data thus opening an exploratory pathway for research. This research builds on the SMOTE-ENC protocol to handle datasets with both nominal and continuous types through ExaSens that advances classification accuracy for COPD detection.

**Identification of Gaps**

Previous literature mostly restricts oversampling methods such as SMOTE and ADASYN to singular datasets containing only continuous or nominal variables. The application of SMOTE and ADASYN remains limited to datasets with either continuous or nominal features while medical datasets which contain both feature types remain underexplored especially in COPD classification studies. Current studies show a gap because medical data analysis has not yet explored the combination of SMOTE-ENC with advanced classification algorithms such as Random Forest or SVM. The models I\_SMOTE and I\_ADASYN produced better recall and F1-score results yet no comprehensive assessments have been conducted on COPD datasets which contain both nominal and continuous data. Our research addresses this existing gap by testing SMOTE-ENC against the ExaSens dataset which should improve classification precision while balancing class distribution especially when detecting medical conditions like COPD.

**Relation to the Research and Hypothesis**

The project presents a novel methodology which utilizes SMOTE-ENC as an advanced oversampling technique that targets mixed-feature datasets thereby addressing a deficiency in existing research approaches. This research builds on current understanding by showing effective performance gains through this new method when applied to medical datasets that feature both categorical and continuous variable types. Through evaluation of the ExaSens dataset which serves as a vital resource for COPD detection this research supports predictive modeling improvement goals in healthcare and sheds light on advanced oversampling methods for medical diagnostics.

**References**

Araf, I., Idri, A. and Chairi, I., 2024. Cost-sensitive learning for imbalanced medical data: a review. *Artificial Intelligence Review*, *57*(4), p.80.

Chachoui, Y., Azizi, N., Hotte, R. and Bensebaa, T., 2024. Enhancing algorithmic assessment in education: Equi-fused-data-based SMOTE for balanced learning. *Computers and Education: Artificial Intelligence*, *6*, p.100222.

Chowdhury, M.M., Ayon, R.S. and Hossain, M.S., 2024. An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset. *Healthcare Analytics*, *5*, p.100297.

Cu, N.G., Nghiem, T.L., Ngo, T.H., Nguyen, M.T.L. and Phung, H.Q., 2024. Increment of Academic Performance Prediction of At‐Risk Student by Dealing with Data Imbalance Problem. *Applied Computational Intelligence and Soft Computing*, *2024*(1), p.4795606.

Firat Atay, F., Yagin, F.H., Colak, C., Elkiran, E.T., Mansuri, N., Ahmad, F. and Ardigò, L.P., 2024. A hybrid machine learning model combining association rule mining and classification algorithms to predict differentiated thyroid cancer recurrence. *Frontiers in Medicine*, *11*, p.1461372.

Glyde, H.M.M., 2024. *Predictive Modelling Acute Exacerbations of Chronic Obstructive Pulmonary Disease* (Doctoral dissertation, University of Bristol).

Hamida, S., Lamrani, D., El Gannour, O., Saleh, S. and Cherradi, B., 2024. Toward enhanced skin disease classification using a hybrid RF-DNN system leveraging data balancing and augmentation techniques. *Bulletin of Electrical Engineering and Informatics*, *13*(1), pp.538-547.

Mukherjee, M. and Khushi, M., 2021. SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features. *Applied system innovation*, *4*(1), p.18.

Rofik, R., Aulia, R., Musaadah, K., Ardyani, S.S.F. and Hakim, A.A., 2024. The Optimization of Credit Scoring Model Using Stacking Ensemble Learning and Oversampling Techniques. *Journal of Information System Exploration and Research*, *2*(1).

Yang, X., 2024. Application and Prospects of Artificial Intelligence Technology in Early Screening of Chronic Obstructive Pulmonary Disease at Primary Healthcare Institutions in China. *International Journal of Chronic Obstructive Pulmonary Disease*, pp.1061-1067.

Zeng, Q., Gong, Z., Wu, S., Zhuang, C. and Li, S., 2024. Measuring cyclists’ subjective perceptions of the street riding environment using K-means SMOTE-RF model and street view imagery. *International Journal of Applied Earth Observation and Geoinformation*, *128*, p.103739.

Zhang, Y., Deng, L. and Wei, B., 2024. Imbalanced Data Classification Based on Improved Random-SMOTE and Feature Standard Deviation. *Mathematics*, *12*(11), p.1709.