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

**Chapter-3**

**Methodology**

The methodology chapter presents the project workflow which utilizes the SMOTE and SMOTE-ENC technique to solve class distribution issues in the ExaSens dataset. The classification methodology achieves COPD model integrity through implementation of machine learning algorithms together with ethical guidelines and evaluation standards.

**Project Design**

The project tackles the class imbalance problem in the ExaSens COPD patient saliva sample dataset using SMOTE as well as SMOTE-ENC. According to Mukherjee and Khushi (2021) this oversampling method helps balance mixed nominal and continuous datasets by creating additional minority class samples to improve COPD classification accuracy. The ExaSens dataset proves to be a suitable test case because it contains multiple data types like smoking status as category and saliva permittivity as continuous measurement. Data collection and preprocessing start the project workflow through missing value handling and encoding nominal features followed by scaling continuous data. The SMOTE and SMOTE-ENC algorithm synthesizes new minority class examples to correct dataset imbalance. Decision Trees alongside SVM and Random Forest models are trained using both original datasets and balanced sets created through oversampling techniques. Rigorous evaluation of these models through metrics including accuracy, precision, recall, F1-Score, AUC-PR, MCC and balanced accuracy reveals how well oversampling techniques supports minority class prediction improvements for COPD conditions. Privacy and anonymity as ethical concerns received attention prior to dataset collection and analysis. The existing public status of the anonymized dataset meant ethical approval was unnecessary.

**A diagram of data preparation

Description automatically generated**

**Figure Project Architecture**

**Dataset Description**

The ExaSens dataset available at IEEE DataPort features valuable information from saliva samples of numerous respiratory disease patients and healthy control subjects. It comprises four sample groups: (I) COPD patients without acute respiratory infection, (II) asthma patients without respiratory infections, (III) individuals with respiratory infections but without COPD or asthma, and (IV) healthy controls. The dataset possesses essential fields such as patient ID alongside demographic information (age and gender), smoking status which distinguishes between non-smokers, ex-smokers and active-smokers and various saliva permittivity measurements including both real and imaginary parts with their minimum and average values. The dataset displays nominal and continuous features which allow researchers to test advanced machine learning approaches for diagnosing respiratory conditions using non-invasive saliva analysis.

**Justification of Dataset**

This research required the ExaSens dataset because of its distinct combination of continuous and categorical features that matched the needs of oversampling process developed for datasets containing mixed data types. This dataset examines COPD which stands as a significant health concern worldwide and reveals its potential to improve diagnostic methods for lesser-known conditions. Considering research that has shown that the medical data is imbalanced and creates biased predictions with reduced model performance (Cu *et al.,* 2024; Firat Atay *et al.,* 2024), there is a need for the application of SMOTE and SMOTE-ENC to correct class imbalance on the ExaSens dataset. This dataset contains diverse features that proved useful to model development, confirming the results of Hamida *et al.,* (2024) and Zhang *et al.,* (2024) that demonstrate improved model performance through increased dataset variety. As the dataset is openly accessible, it encourages reproducibility and fosters further research activities in medical diagnostics as seen in Rofik *et al.,* (2024).

**Choice of Methods**

SMOTE and SMOTE-ENC are used as oversampling techniques to tackle class imbalance for COPD classification in the ExaSens dataset in this research. Prior studies that show the effectiveness of SMOTE variations when handling imbalanced datasets guide the selection of these methods. In credit assessment models (Rofik *et al.,* 2024), SMOTE has shown improvements in recall and F1-score, which makes it a good model in medical data where the detection of minority class is important. Furthermore, the capability of SMOTE-NC to deal with mixed categorical and continuous data, was shown in DTC recurrence prediction (Firat Atay *et al.,* 2024), which matches the structure of the ExaSens dataset.

According to the methodology proposed by Mukherjee and Khushi (2021), SMOTE-ENC is chosen because it specifically handles the shortcoming of the standard SMOTE in datasets with nominal as well as continuous features. One of the ways in which the aforementioned studies further supports the importance of balancing techniques is through Equi-Fused-Data-based SMOTE (Chachoui *et al.,* 2024) and FSDR-SMOTE (Zhang, Deng, and Wei, 2024) which highlight the need in better oversampling to gain improved classification performance. DT, SVM as well as RF are utilised as ML models, since these models have been effective as found in previous research on hybrid classification and ensemble learning (Hamida *et al.,* 2024; Zeng *et al.,* 2024). The performance evaluation is conducted (by metrics such as accuracy, precision, recall, F1-score, AUC-PR, MCC and balanced accuracy) to assess the effect of oversampling techniques on classification results.

**Justification and Support of Choices**

Recent studies stress the importance of SMOTE based methods to alleviate class imbalance in medical datasets, hence SMOTE and SMOTE-ENC were selected as the oversampling techniques for COPD classification. Trigka and Dritsas’ (2025) study showed that an improved SMOTE approach indeed increases model performance in terms of recall and AUC-ROC, and thus, the use of SMOTE variants to create more realistic synthetic samples is crucial. Furthermore, according to Cheng *et al.,* (2025) and Zuhria and Habibi (2025), SVM as well as RF are also efficient for medical classification task and RF performs well in terms of classification speed whereas SVM is best in terms of precision. Given the mixed nominal and continuous attributes of the ExaSens dataset, SMOTE-ENC was chosen due to its effectiveness on such datasets, similar to its use in federated learning experiments of Cheng et al. (2025). Moreover, Maheshwari *et al.,* (2025) also state that DT can be used to classify the medical transactions, which corroborates its suitability in classifying COPD. Trigka and Dritsas (2025) results show that hybrid and enhanced oversampling techniques for class imbalance are justified by the improvements they provide, leading to SMOTE-ENC being included as a method to represent minority class distributions. These studies together show that SMOTE, SMOTE-ENC, RF, and SVM are chosen as good methods to promote classification performance in COPD prediction by ExaSens dataset.

**Use of Tools and Techniques**

Throughout the duration of the project, Python was used as the main programming language since it had a great support for machine learning, and data analysis. Machine learning models were installed and their performance was also measured utilizing Scikit learn together with Imbalanced learn for the process of oversampling and using Matplotlib and Seaborn for data visualization and analysis of results. NumPy and Pandas tools were used in data preprocessing activities and data manipulation tasks. For the project to be successful, the tools used must be able to provide a dependable functionality and good documentation that also meet project needs. During the comparative evaluation, available libraries were proven capable to work properly with mixed–feature datasets, giving the research a possibility to efficiently deploy oversampling methods and machine learning workflows.

**Introduction to Sampling Techniques**

Sampling methods are heavily used for class imbalance management, and they are used to ensure that minority class instances are adequately represented. In this research SMOTE and SMOTE-ENC methods were chosen because it generates synthetic samples for datasets that mix nominal and continuous features to overcome the imbalance efficiently.

**SMOTE**

The ExaSens dataset achieves balanced class distribution by generating synthetic samples for the minority class by interpolation of existing data points using SMOTE (Trigka and Dritsas, 2025). Creating the new sample as:

Where -minority class instance

- nearest neighbor

- random value between 0 and 1

**SMOTE-ENC Technique**

The SMOTE-ENC algorithm functions as an advanced oversampling solution designed specifically for datasets which incorporate both nominal and continuous attributes. The original SMOTE algorithms perform adequately on continuous data but do not handle nominal features effectively since interpolation operations lose meaning when applied to categorical data. Through SMOTE-ENC extension SMOTE transforms nominal data into integer representations which permits the algorithm to perform interpolation techniques on continuous features while managing nominal data correctly. To achieve class- balance the algorithm creates synthetic data points within the minority class feature space (Mukherjee and Khushi, 2021).

The process involves two primary steps:

**Nominal Feature Handling:** The representation of nominal features uses integer values for encoding purposes. The synthetic data generation procedure chooses nominal values for new samples randomly from the corresponding feature values of KNN which belong to the minority class.

**Continuous Feature Interpolation:** Synthetic examples for continuous features result from interpolation between a minority class sample and one of its K-NN. This is achieved using the following formula:

**New­­\_ Sample =**

Where:

-minority class instance, - One of the K-NN of

- random number in the range [0, 1]

The SMOTE-ENC technique merges multiple strategies to protect nominal features and produce realistic synthetic examples of continuous features which leads to balanced data readiness for machine learning models.

**Algorithm Steps**

* Determine minority class data points and identify their k closest data neighbors in the dataset.
* Every minority instance 𝑋𝑖 requires selection of one neighbor from its k nearest neighbors 𝑋𝑗.
* The formula above guides interpolation which generates new values for the continuous features.
* For nominal features, randomly select values from 𝑋𝑖 or 𝑋𝑗.
* Keep repeating the procedure until the dataset reaches the required oversampling measurement.

The generation technique produces representative samples from original data that handle mixed data types thus proving SMOTE-ENC to be highly effective for COPD classification tasks within the ExaSens dataset.

A diagram of a process

Description automatically generated

**Figure SMOTE-ENC Algorithm**

**Algorithms Used in the Research:**

In this research, Decision Trees along with SVM and Random Forest classifiers are used to evaluate model performance using datasets that remain unbalanced and those balanced by oversampling methods. The researchers chose these algorithms because they show a successful result in classification tasks and they can work with complex datasets.

**Decision Tree (DT):**

The datasets are divided hierarchically by the algorithm until the prediction result is reached using variable values (Mujahid *et al.,* 2024). The splitting criterion for the node separation process as shown in figure is either Gini Index or Information Gain.

Formula for Gini Index:

**Where – proportion of class i**

A diagram of a tree

Description automatically generated

**Figure Decision Tree Model**

The algorithm can serve as a simple tool for the determination of the important characteristics influencing the grouping of COPD taking into account reliable managing data in various forms.

**Support Vector Machine (SVM)**

The best hyperplane that can separate classes is calculated by this classification algorithm that extends the spacing between the classes (Gomiasti *et al.,* 2024), and the kernel trick is used to manage nonlinear data.

Formula for Hyperplane

Where is the weight vector and b is the bias

In particular, the technique supports processing of high dimensional data and retains high performance when determining decision boundaries in the very same complex data such as ExaSens.

**Random Forest (RF):**

This is an ensemble learning technique, which is based on the creation of multiple decision trees as shown in figure, and the output is the combination result through majority voting (Nizam‐Ozogur and Orman, 2024). The ensemble approach of the model works on balanced and imbalanced data and provides better prediction accuracy along with lowering overfitting.

Formula for prediction

Where is the prediction from 1st  Tree

A diagram of a tree

Description automatically generated

**Figure Random Forest Model**

**Performance Evaluation**

The metrics used for model evaluation include accuracy, precision, recall, F1 score, AUC-PR, MCC, and balanced accuracy. Complete performance assessments for the classifiers on the datasets are delivered by the evaluation metrics, irrespective of the balance of the datasets. Research from Chowdhury, Ayon, Hossain (2024) together with Zhang, Deng, Wei (2024) shows why multiple metrics are essential for evaluating how oversampling techniques affect model performance. The measurement of precision involves calculating true positive predictions as part of the entire predicted positive population which shows how well a model detects instances of the minority class. The recall metric serves as sensitivity by measuring the degree to which true positive results match actual positive cases to assess model performance for minority classes. In imbalanced dataset situations the F1-Score provides utility through its harmonic mean which combines both precision and recall measurement metrics. The AUC-PR metric demonstrates the balance between precision and recall along different thresholds to analyze model effectiveness particularly for minority class elements (Gomiasti *et al.,* 2024). The MCC metric establishes its correlation between predicted classifications and actual outcomes through analysis of both true/false positives and negatives which enhances its effectiveness with imbalanced data sets (Nizam‐Ozogur and Orman, 2024). Balanced accuracy derives its value through averaging recall rates across all classes so evaluation remains fair to both major and minor classes. The set of metrics conducts a complete assessment of classification models by balancing prediction outcomes and managing class imbalances for COPD results.

A math equations on a white background

Description automatically generated

**Figure Confusion Matrix and Evaluation metrics formula (Nizam‐Ozogur and Orman, 2024)**

**Test Strategy**

The testing strategy designed for this project used several layers to verify both system reliability and accurate results. Testing of the preprocessing pipeline together with oversampling implementation and machine learning models through unit testing verified their individual functionality. Through integration testing the project verified smooth operational links between the balanced datasets and classifiers while system testing confirmed each sequential step from data preprocessing through performance assessment functions correctly. The performance testing measures SMOTE as well as SMOTE-ENC computational efficiency alongside its effects on model training duration. The system underwent extensive evaluation which provided complete reliability through a multi-level approach.

**Testing and Results**

The evaluation process tested machine learning algorithms like Decision Trees, SVM and Random Forest against both original datasets and balanced datasets. The performance was evaluated using several effectiveness indicators such as accuracy, precision, recall, F1-Score, AUC-PR, MCC and balanced accuracy. The applications of SMOTE and SMOTE-ENC yielded significantly better predictions on the minority class and better recall and F1-Score values for dealing with imbalanced datasets. When the project visualized the results of the comparative analysis, it was evident that the oversampling techniques had benefits, as charts and confusion matrices were used.

**Validation**

Multiple quality assurance methods were established to provide guaranteed proof of outcome precision and dependability. The performance of the models on the new data was tested with cross validation. The improvements from using oversampling techniques were demonstrated against the results from the original dataset. AUC-PR and MCC metrics are used for the evaluation of the models as they consider both false negatives as well as positives, and the models were shown to be robust in these metrics. It ran an error analysis on misclassified cases in order for identifying patterns that need to be addressed. Stable results validating the research objectives were achieved through the collective application of validation techniques.

**Ethical Issues**

Examination of the ExaSens dataset qualifies for exemption from ethics review because it is publicly accessible and maintains anonymity by removing all personal identifiers. The analysis respects privacy protection throughout the research by following accepted ethical data processing requirements. SMOTE and SMOTE-ENC achieves dataset balance while maintaining its original data structure which helps prevent both bias and potential harm. Multiple protective measures which include data security efforts and adherence to ethical norms guarantee secure treatment of sensitive medical data during this entire project.

**Legal Issues**

The public dataset used in the research does not contain personal information, so that the data protection rules such as the GDPR are met. This research objective fits perfectly with the utilization of the dataset and its license permits its use for research purposes. There is strict security procedure in place to ensure that unauthorized access is prevented to the data storage area and processing area so as to prevent security breaches. Once the results of the project are released, they will be done so according to legal requirements, with respect given to data confidentiality and intellectual property rights. In the current dataset, the data is anonymized and live data collection requires ethical approval and informed consent as well as good data security to avoid misuse and unauthorized access.

**Professional Issues**

The project has a structured methodology which is associated with having clear objectives that keep each stage accountable to maintain quality. The best practices for preprocessing and measuring model performance against the benchmarks are used with the machine learning models and oversampling implementation. The data processing and result presentation procedure meets established academic and professional criteria for every procedure. The project achieves professional excellence by validating results deeply, and consequently tackling limitations, to establish scalable and reproducible foundations for future work.

**Social Issues**

This project seeks to improve healthcare outcomes in general for underrepresented patient populations through improved COPD condition prediction methods. Data generation in a balanced manner leads to improved equitable performance of the machine learning systems as they can detect patterns in the minority group more accurately. This project also addresses societal risks, diagnosis bias, by means of improvements in model generalizability. With saliva samples, there are non-invasive diagnostics while keeping accessibility so that healthcare services lead to equal outcomes for real world users.

**Practicality**

This project was implemented by solving multiple challenges, focusing on data preprocessing for mixed type datasets and using SMOTE and SMOTE-ENC to balance class. Owing to its high computational load to meet processing time requirements, the oversampling of large datasets required optimization strategies. Decision Trees and Random Forests were preferable for mixed type data and SVMs were good at handling class imbalance. The performance metrics evaluation was difficult due to the fact that testing involved both balanced and unbalanced datasets, making the results hard to interpret. By analyzing relevant parts of the dataset selectively, this work solved time limitations and computational capacity issues by using cloud computing scalability. Unfortunately, the approach encountered many obstacles, but it also proved to work well in relation to project goals and produce results that are easily applicable for a future real world COPD diagnostic task.

**Conclusion Summary**

The chapter offered an extensive review of several techniques which covered data preprocessing steps and oversampling implementation together with various balancing and classification algorithms. The subsequent chapter will explore experimentation through which the presented methods undergo testing and detailed analysis of their outcomes.

**References**

Mujahid, M., Kına, E.R.O.L., Rustam, F., Villar, M.G., Alvarado, E.S., De La Torre Diez, I. and Ashraf, I., 2024. Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. *Journal of Big Data*, *11*(1), p.87.

Gomiasti, F.S., Warto, W., Kartikadarma, E., Gondohanindijo, J. and Setiadi, D.R.I.M., 2024. Enhancing lung cancer classification effectiveness through hyperparameter-tuned support vector machine. *Journal of Computing Theories and Applications*, *1*(4), pp.396-406.

Nizam‐Ozogur, H. and Orman, Z., 2024. A heuristic‐based hybrid sampling method using a combination of SMOTE and ENN for imbalanced health data. *Expert Systems*, p.e13596.

Cheng, A., Zhang, J., Sharma, A., Gupta, V. and Guo, Y., 2025. A Systematic Comparison of Horizontal Federated Learning Algorithm Based on Random Forests in a Medical Setting. *Machine Intelligence Research*, pp.1-13.

Zuhria, L. and Habibi, A.R., 2025. Comparative Analysis of Random Forest and SVM Performance in Asthma Prediction. *Sinkron: jurnal dan penelitian teknik informatika*, *9*(1), pp.347-356.

Maheshwari, S., Jain, P.K., Al‐Sharify, N.T., Ghosh, S., Dubey, D. and Kaur, G., 2025. Medical data analysis and transaction type prediction using machine learning and blockchain technology. *IET Blockchain*, *5*(1), p.e70001.

Trigka, M. and Dritsas, E., 2025. Improving Cardiovascular Disease Prediction with Deep Learning and Correlation-aware SMOTE. *IEEE Access*.

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.

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.

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.

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).

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.

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.