Beyond Accuracy: Evaluating Sampling Techniques and Cost Sensitive Learning for Class Imbalance in Healthcare Data

1st Ananta Dutta Choudhury
Dept. of Computer Science
Jamia Millia Islamia
New Delhi, India
ananta2407471@st.jmi.ac.in

2nd Siddharth Sharavat
Dept. of Computer Science
Jamia Millia Islamia
New Delhi, India
sharavatsiddharth@gmail.com

3rd Mohd Amir Pasha
Dept. of Computer Science
Jamia Millia Islamia
New Delhi, India
amirpashajan7@gmail.com

***Abstract*— The performance of any model on a dataset can be hindered insidiously by class imbalance, especially with medical data, where the majority class is usually the "healthy" kind and minority class, the "diseased" kind. In these cases, it becomes more important to identify the minority class samples and avoid False Negatives to prevent misdiagnosing the patients who need help. The Centers for Disease Control and Prevention (CDC) Diabetes Health Indicators Dataset is a good fit to explore such a problem due to its large class imbalance issue.**

We explore the effectiveness of various resampling techniques such as Random Oversampling (ROS), Random Under sampling (RUS), SMOTE, Borderline-SMOTE, ADASYN, SMOTE-ENN, and SMOTE-Tomek on the dataset. We also experiment with Cost Sensitive learning and apply the CatBoost classifier, optimized with hyperparameter tuning. The performance of each resampling method was studied using various evaluation metrics like Precision, Recall, ROC-AUC etc. across stratified 5-fold cross-validation. Results indicate that while resampling techniques clubbed with Cost Effective learning improved recall, the original dataset with class weights provided a competitive balance across all metrics. These findings highlight the trade-off between Recall and overall performance when addressing class imbalance in healthcare data.

Keywords—*Class Imbalance, Cost Effective Learning, Resampling techniques, SMOTE, Medical, Health, Diabetes*

I. Introduction

As of this date, an estimated 828 million people globally are living with diabetes, with over 450 million people being unaware of their condition, and the number is projected to grow. Early and correct detection of this disease can prevent complications in the near future. Medical data, such as the Centers for Disease Control and Prevention (CDC) Diabetes Health Indicators Dataset [1], is a popular choice among researchers for identifying and preventing diabetes using Machine learning (ML). However, the challenge with medical data is that the number of people suffering from the disease is always low compared to the number of people who are healthy. ML models trained on such highly imbalanced data tend to deliver high accuracy but report low recall as it is easy for a model to just report "healthy" when majority samples are already of the "healthy" class. This leads to a lot of False Negatives (In this case, predicting a "diabetic" person as "healthy") which can lead to detrimental implications, including delay in treatment, increase in financial burden and several other complications.

To tackle class imbalance, we explore various Resampling techniques like Random Oversampling (ROS), which creates copies of underrepresented class samples, and Random Undersampling (RUS), which drops overrepresented class samples randomly. We further employed Synthetic Minority Oversampling Technique (SMOTE) [2], using which synthetic samples are generated for the minority class by interpolating between existing underrepresented data points. In addition, we tested Adaptive Synthetic Sampling (ADASYN) [3], a variant of SMOTE that focuses on augmenting more samples in regions where the underrepresented class is harder to learn. To explore more sophisticated hybrid methods, we applied Borderline-SMOTE (types 1 and 2) [4], which specifically generates synthetic samples near the decision boundary between classes, SMOTE combined with Edited Nearest Neighbors (SMOTE-ENN) [5], which removes noisy samples after oversampling, and SMOTE combined with Tomek Links (SMOTE-Tomek) [5] which eliminates overlapping samples from all classes after oversampling.

We then experiment with Cost Effective Learning (CEL) [6] using class weights achieved from CatBoost using RandomizedSearchCV. The motivation behind using CEL lies in its ability to handle class imbalance while minimizing the overall misclassification cost, which is particularly critical in medical datasets where false negatives can be more harmful than false positives. These approaches are elaborately explained in the upcoming sections, and the results and trade-off are discussed to identify the right strategy according to the use case and aim.

II. Related Work and Literature Review

Several studies have explored the use of the SMOTE [2] to improve predictive performance on imbalanced health data. For eg. Ramezankhani et al. [6] evaluated SMOTE on the Tehran Lipid and Glucose Study cohort and observed significant improvements in sensitivity for various classifiers including probabilistic neural networks and naive Bayes models albeit with a few trade-offs in accuracy. Similarly, Wibowo et al. [7] applied SMOTE to an Indonesian hospital's data and discovered that SVM's with SMOTE demonstrated an improved recall compared to models without the use of oversampling.

Hybrid methods that combine both over and undersampling are becoming more popular. One notable example applied SMOTE followed by SMOTE-ENN [5] for detecting missed abortions and diabetes cases, thereby significantly enhancing classification performance over SMOTE alone. These hybrid strategies like SMOTE with Tomek Links [5] or SMOTE-ENN, aim to augment the minority class and remove unwanted noise from the majority class to improve the model generalization.

More recent research also explores advanced oversampling variants built for medical data. For example, Praveenkumar & Gunasundari [8] suggested an H-SMOTE tree which integrates Hoeffding Adaptive Trees and SMOTE to refine prediction. Combining SMOTE with ensemble learning methods has provided promising results in disease prediction. A study using SMOTE-

augmented datasets with AdaBoost and XGBoost ensembles achieved high AUC scores (around 0.968) [9] for diabetes detection, pointing to the complementary strength of resampling and ensemble methodology.

Furthermore, imbalanced data resampling has been applied in various medical contexts ranging from cardiovascular issues to chronic kidney disease and pediatric diagnoses. Techniques like clustering-based undersampling [10] and hybrid SMOTE-ENN strategies [11] have shown superior performance over standard methods in large-scale medical classification tasks.

However, despite these advances, several limitations persist. Oversampling methods like SMOTE may lead to overfitting by generating synthetic samples that are too similar to existing minority class instances. Hybrid approaches, while more effective, often increase computational complexity and may inadvertently remove informative samples during the noise-reduction step. Ensemble-based methods, though powerful, require careful hyperparameter tuning and can become resource-intensive, limiting their scalability in real-world medical applications. Moreover, a critical gap lies in the limited focus on minimizing the cost of misclassification. In medical datasets, false negatives (missed diagnoses) are often far more harmful than false positives, yet many studies emphasize overall accuracy or AUC without explicitly addressing this asymmetry.

To address these gaps, our study not only compares basic resampling techniques and variations of SMOTE to evaluate their robustness but also explores Cost Effective Learning (CEL). By incorporating class weights derived from CatBoost hyperparameter tuning, we aim to skew the model towards improved recall and clinically meaningful performance, thereby overcoming the limitations of purely accuracy-driven resampling approaches.

III. METHODOLOGY

The methodology of this study is proposed in Fig.1. The dataset is preprocessed and then onwards, we bifurcate our study into two approaches (based on Cost Effective Learning). Each approach gets a copy of the preprocessed dataset and then the resampling methods are applied and passed on to the model. The results are then studied and discussed.

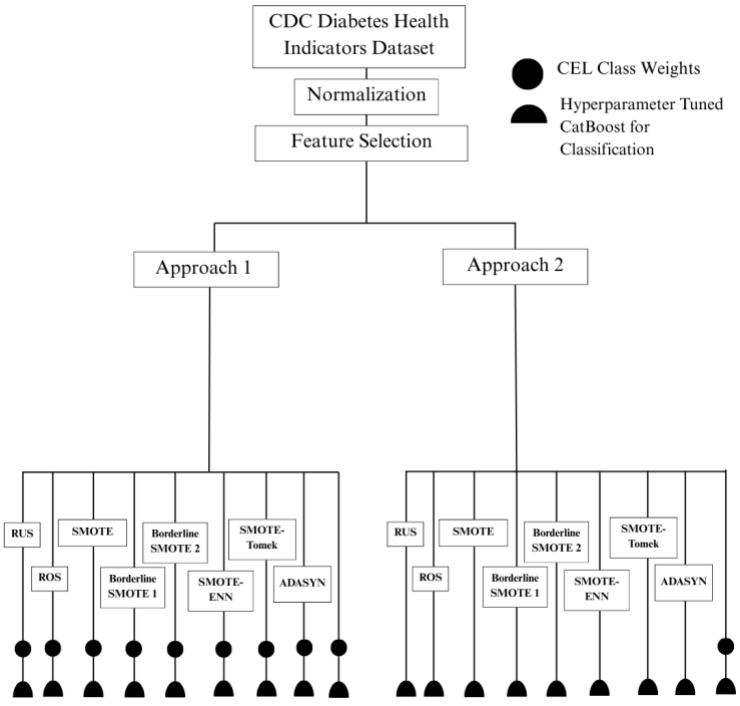

Fig.1. Methodology of study

1. Dataset

For this study, we used the 2014 Behavioral Risk Factor Surveillance System (BRFSS) dataset provided by the CDC [1] comprising of 21 attributes.
The target column $y \in \{0,1\}$. If the patient is diabetic (called the positive class here), y=1, else y=0 (the negative class).

The imbalance ratio measured (as shown in Eq. 1) is quite high, i.e. 6.18 (Fig. 2). [13]

$$Imbalance\ Ratio = \frac{Majority\ class\ samples}{Minority\ class\ samples} \quad (1)$$

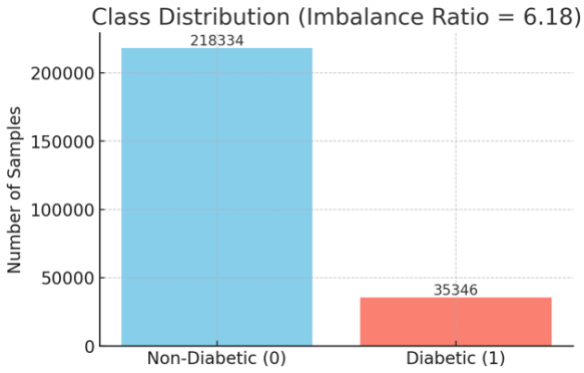

Fig.2. Class Imbalance in dataset

2. Data Preprocessing and Model Selection

In this study, we are emphasizing on understanding and comparing the effect of resampling techniques on healthcare data. According to Xinyi [13] CatBoost has shown to perform the best compared to ML models like Random Forest, Decision tree, Gradient Boosting, Gaussian Naive bayes, logistic regression and linear discriminant. CatBoost, a gradient boosting algorithm based on decision trees, is particularly effective for tabular datasets. It incorporates built-in handling of categorical features, employs a unique boosting approach to limit overfitting, and demonstrates strong performance even under class imbalance conditions [14]. Since the focus of our study is on tackling class imbalance, a singular model, i.e. CatBoost is used to conduct our experiments.

2.1. Normalization

Continuous variables are standardized using z-score normalization.

$$z = \frac{x-\mu}{\sigma} \quad (2)$$

Since the range of these attributes are quite different, normalization is important to ensure one attribute doesn't overpower the other.

2.2. Feature Selection

The dataset contains 21 attributes and not all of them contribute significantly to the classification of diabetic patients. We applied CatBoost to the data and extracted the most relevant features using its feature importance method.

The top eleven features were selected, namely, "GenHlth", "Age", "BMI", "HighBP", "HighChol", “CholCheck”, "HvyAlcoholConsump", "Sex","Income", ”HeartDiseaseorAttack”, ”DiffWalk” (Refer to Fig.3. and Table 1.)

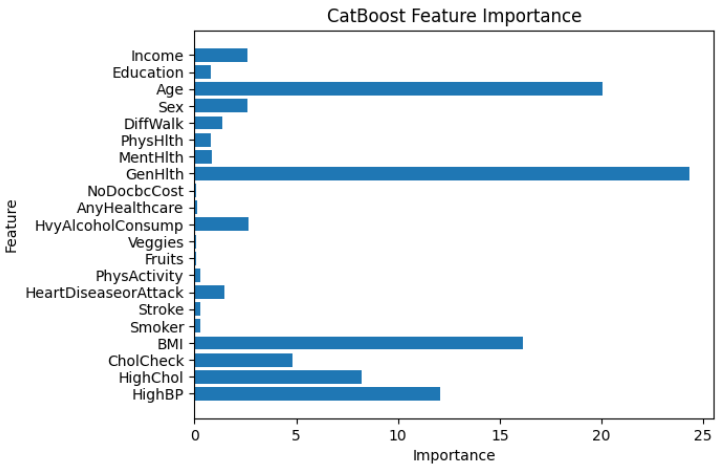

Fig.3. Feature Importance using CatBoost

Index	Feature	Importance
1	GenHlth	24.31641747
2	Age	20.08667134
3	BMI	16.14666168
4	HighBP	12.05470394
5	HighChol	8.224571414
6	CholCheck	4.795719274
7	HvyAlcoholConsump	2.648914787
8	Sex	2.589351739
9	Income	2.3506095
10	HeartDiseaseorAttack	1.456900122
11	DiffWalk	1.395907202
12	MentHlth	0.866478185
13	Education	0.808452523
14	PhysHlth	0.784455041
15	PhysActivity	0.72465004
16	Stroke	0.28889541
17	Smoker	0.14969093
18	AnyHealthcare	0.116590073
19	Fruits	0.0919932
20	NoDocbcCost	0.087916382
21	Veggies	0.08636832

TABLE 1. Features with Decreasing Importance

2.3. Handling Class Imbalance

Due to uneven class samples in the dataset, we now apply various Resampling techniques (only on the training folds of the 5-fold cross validation, to prevent data leakage) [14].

2.3.1. ***Random Oversampling (ROS)****:* Generates randomly duplicates of minority class samples to balance the dataset. Although, it can lead to overfitting as the model trains on duplicate data.

2.3.2. ***Random Undersampling (RUS)***: Randomly discards samples from the majority class until the size of both classes become equal. However, this can lead to potential loss of information.

2.3.3. ***SMOTE (Synthetic Minority Oversampling Technique)***
It generates new synthetic samples by interpolating between existing minority class samples using k-nearest neighbors [2] as follows:

$$x_{\{new\}} = x_i + \delta.(x_{\{nn\}} - x_i) \quad (3)$$

Where:
x_i= a minority class sample
$x_{\{nn\}} = one\ of\ its\ nearest\ neighbors\ (minority\ class)$
$\delta = a\ constant\ between\ 0\ and\ 1$

However, the synthetic data generated might not follow your original data’s distribution. Furthermore, it has trouble handling categorical data (CDC diabetes dataset [1] has a lot of them) and it is quite sensitive to outliers.

2.3.4. ***ADASYN (Adaptive Synthetic Sampling)*** is an extension of SMOTE that generates data samples in areas where the classifier finds it hard to classify. Instead of generating samples uniformly, it adapts the number of synthetic samples per minority instance based on difficulty of classification. Minority samples that are harder to learn (surrounded by majority class neighbors) get more synthetic samples [3].

2.3.5. ***Borderline SMOTE (B1 AND B2)*** is an improvement over SMOTE. It focuses only on borderline minority samples (i.e., those near the decision boundary), since they are more at risk of being misclassified.

Borderline-SMOTE1 (B1) generates new synthetic samples only for the borderline minority samples and interpolates between borderline samples and other minority neighbors.

Borderline-SMOTE2 (B2) is like B1 but also allows interpolation with majority neighbors. This creates more aggressive synthetic samples, placing them closer to the decision boundary [4].

2.3.6. ***SMOTE-ENN (SMOTE + Edited Nearest Neighbor)*** uses SMOTE to oversample minority and applies ENN to remove noisy or mislabeled points. It produces cleaner decision regions [5].

2.3.7. ***SMOTE-Tomek*** applies SMOTE to oversample minority classes and removes Tomek links (pairs of nearest neighbors belonging to different classes). This helps in reducing Class Overlap. A Tomek link is described as:

$$\forall k,\ d(x_i, x_j) \le d(x_i, x_k)\ \ and\ \ d(x_i, x_j) \le d(x_j, x_k) \quad (4)$$

Where, $x_j\ and\ x_i$ should belong to different classes.

3. Hyperparameter Optimization of the model

We performed hyperparameter optimization for the CatBoost classifier using RandomizedSearchCV with k-fold stratified cross-validation, where each fold keeps the same class distribution. It tries different combinations of parameters like number of trees, tree depth, learning rate, regularization, subsampling, and class weights. The optimization is guided by the recall score (important for detecting diabetes cases) rather than accuracy because accuracy in this case can be misleading. Say, if only 10% patients have diabetes, a classifier predicting "no diabetes" for everyone gets 90% accuracy but 0% recall. In disease prediction, false negatives are dangerous because undiagnosed patients may not receive treatment timely. We then fit the model on the normalized dataset and note the best parameters. The class weights obtained using this model are used for Cost sensitive learning.

IV. Evaluation Metrics

In healthcare, the priorities shift. Instead of looking at overall performance of a model, we start gauging and putting more emphasis on whether the model can correctly identify diabetic patients (the minority class here). The focus is on reducing False Negatives to a minimum as a "diabetic" flagged as "non-diabetic" could be detrimental. The metrics we are using are:

4.1. Accuracy measures the overall percentage of correctly classified samples (Eq. 5). The issue with imbalanced data is, if 90 percent of patients are non-diabetic, a model that always predicts "non-diabetic" will have 90% accuracy regardless of where it is able to correctly classify the remaining 10% diabetic patients.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+FP} \quad (5)$$

4.2. Precision calculates how many of the predicted positives are correct (Eq. 6). A false positive (FP) means a healthy person flagged as diabetic. This leads to unnecessary tests/treatments and can be an issue when medical testing and prescriptions are expensive/labor intensive.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

4.3. Recall measures how many actual positives (diabetes) the classifier correctly identifies (Eq. 7). In medicine Recall is crucial because a false negative (FN) means a sick patient has been incorrectly labeled as healthy. This would be detrimental as it could lead to delay in treatment, leading to multiple complications.

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

4.4. F1 score is an intriguing way to gauge your model's performance on an imbalanced dataset since it combines both precision and recall (Eq. 8). In [1] most people are non-diabetic. If a model always predicts "non-diabetic," it will get high accuracy but detect zero diabetic patients. If the model just predicts "diabetic" for everyone, recall will be 100%, but precision will be terrible (lots of false alarms). If the model is very strict, it may only predict "diabetic" in rare obvious cases, giving high precision but missing most patients (low recall).

The F1-score reflects both precision and recall, rewarding models that not only catch as many diabetic patients as possible (high recall) but also ensure that those labelled diabetic are truly likely to have diabetes (high precision).

$$F1\ score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (8)$$

V. Experimentation and Result

Our study comprises of two different approaches (Fig.1). *Approach 1* involved using CatBoost optimized with hyperparameter fine tuning and then applying a new sampling technique each time on the original dataset before classification. In addition, we apply class weights (Cost Effective Learning) giving the "diabetic" class more importance than the non-diabetic class. In this approach, we apply Cost Effective Learning on the original dataset as well as the Resampled datasets (essentially doubling down on ensuring that our model returns high Recall values). We apply cross validation to check for the model's performance and hence can be sure that what we have achieved is an objective report.

Method	Accuracy	Precision	Recall
Original	0.75285399	0.32904113	0.74463871
Random Oversampling	0.46784926	0.20356378	0.96800204
Random Undersampling	0.46319379	0.20239311	0.97001075
SMOTE	0.66197178	0.27260834	0.85480677
ADASYN	0.65137575	0.26739102	0.86335087
Borderline-SMOTE1	0.65179754	0.26786997	0.86493521
Borderline-SMOTE2	0.5279486	0.22015702	0.93931421
SMOTEENN	0.75249133	0.32525471	0.72254286
SMOTETomek	0.66130558	0.27228766	0.85545748
Method	**F1-Score**	**ROC-AUC**	**PR-AUC**
Original	0.45640563	0.83011361	0.43286268
Random Oversampling	0.33638768	0.82936658	0.43202237
Random Undersampling	0.33490762	0.82936524	0.43062567
SMOTE	0.41338359	0.82271492	0.4162326
ADASYN	0.40832001	0.82092267	0.41167237
Borderline-SMOTE1	0.40905563	0.82092251	0.40370043
Borderline-SMOTE2	0.3567085	0.81094801	0.37476974
SMOTEENN	0.44857991	0.82262826	0.42220314
SMOTETomek	0.4130907	0.82286119	0.41751117

TABLE II. Evaluation Metrics on Applying Class Weights to All datasets (Resampled + Original Dataset)

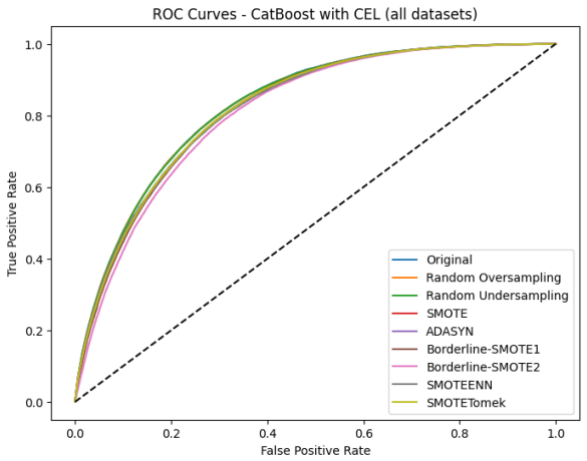

Fig.4. ROC curve for all sampling techniques with CEL on all datasets

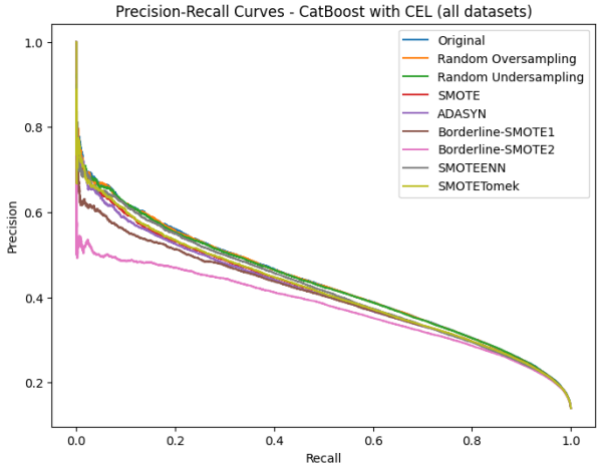

Fig.5. PR curve for all sampling techniques with CEL on all datasets

The evaluation of resampling techniques with Cost-Effective Learning (CEL) on the CDC resampled datasets shows clear trade-offs: Random Oversampling, Random Undersampling, SMOTE, and ADASYN achieved very high recall (≥0.86) but suffered from low precision and accuracy, making them less balanced. Borderline-SMOTE1 and Borderline-SMOTE2 pushed recall even higher (up to 0.93) but further reduced precision, increasing false positives. In contrast, SMOTE-ENN provided the best overall balance, with the highest F1-score (0.448), competitive recall (0.72), and stable ROC-AUC (0.82). The Original dataset with CEL also performed strongly, demonstrating the effectiveness of cost-sensitive learning without resampling. Overall, SMOTE-ENN (with CEL applied to it) is the most balanced choice, while Borderline-SMOTE2 (with CEL) is ideal when maximizing recall is the priority.

Approach 2 involved using the same model with hyperparameter tuning but, instead of applying Cost Effective Learning to all the resampled datasets, we apply it only on the original Dataset. This is to prevent the double emphasis we put on minority samples in Approach 1 (First by resampling it, then by giving it a higher weight than the majority samples).

Method	Accuracy	Precision	Recall
Original	0.75285399	0.32904113	0.74463871
Random Oversampling	0.72120782	0.30680428	0.79474339
Random Undersampling	0.71764822	0.30459838	0.80003395
SMOTE	0.84761905	0.44564146	0.38386239
ADASYN	0.85028382	0.45231371	0.35342047
Borderline-SMOTE1	0.84438663	0.43473864	0.38918124
Borderline-SMOTE2	0.78584043	0.34433328	0.59395688
SMOTEENN	0.83401529	0.41925236	0.4965767
SMOTETomek	0.84698439	0.44481892	0.39580151
Method	**F1-Score**	**ROC-AUC**	**PR-AUC**
Original	0.45640563	0.83011361	0.43286268
Random Oversampling	0.44270562	0.82966522	0.43246933
Random Undersampling	0.44121295	0.82922488	0.42774708
SMOTE	0.41245136	0.82176477	0.41204553
ADASYN	0.39679817	0.81999372	0.40662673
Borderline-SMOTE1	0.41070042	0.81966673	0.39636468
Borderline-SMOTE2	0.43594003	0.80593141	0.35723325
SMOTEENN	0.45465024	0.8243034	0.42514181
SMOTETomek	0.41888109	0.8221075	0.4135456

TABLE III. Evaluation Metrics on Applying Class Weights to Only Original Dataset

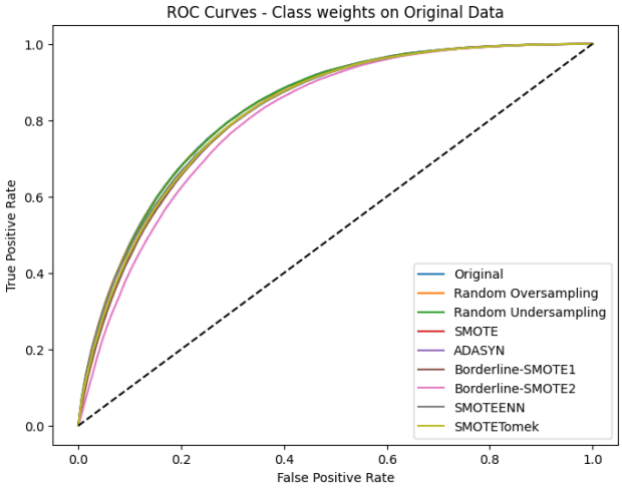

Fig.6. ROC curve for all sampling techniques with CEL only on original dataset

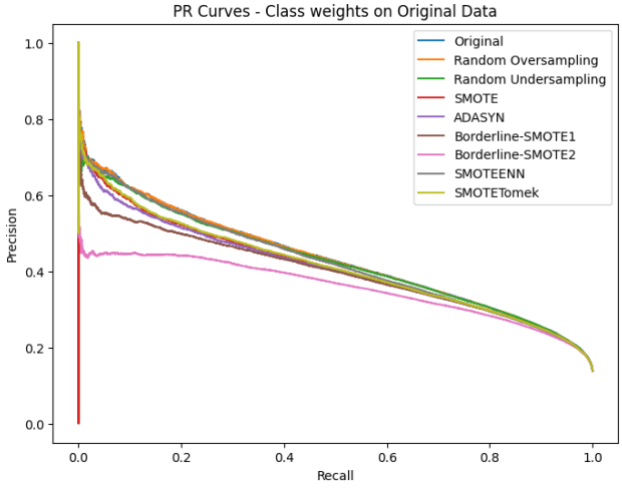

Fig.7. PR curve for all sampling techniques with CEL only on original dataset

When applying class weights only to the original dataset, the model achieved strong recall (0.7446) and ROC-AUC (0.8301), showing that cost-sensitive learning alone significantly boosts minority class detection. Random Oversampling further improved recall (0.7974) but lowered precision, while Random Undersampling also showed decent recall (0.8000) but struggled with precision. SMOTE and its variants offered balanced improvements, with SMOTE achieving the best overall accuracy (0.8476) and ADASYN reaching high recall (0.8635) though at the cost of precision. Borderline-SMOTE methods improved recall moderately, whereas SMOTE-ENN provided the most balanced F1-score (0.4547). Overall, cost-sensitive learning on the original dataset proved effective in enhancing recall, but oversampling methods like ADASYN and SMOTE showed complementary benefits depending on whether higher recall or balanced performance was prioritized.

VI. DISCUSSION

Approach 1 involved applying resampling techniques to the dataset and additionally using class weights to penalize misclassification of the "diabetic" class. *Approach 2* involved using class weights only on the original dataset and not on the resampled datasets.

The experimental results indicate that Random Oversampling (ROS) and ADASYN delivered strong recall performance, particularly when paired with Cost Effective Learning (CEL). This suggests that these techniques were effective in amplifying the detection of diabetic class samples (minority). On the other hand, SMOTE and its variants (Borderline-SMOTE1/2, SMOTE-ENN, SMOTE-Tomek) produced moderate but consistent gains. These methods improved class balance by generating synthetic samples, though their impact varied: SMOTE-ENN offered a more balanced enhancement in F1-score, Borderline-SMOTE1/2 boosted recall but occasionally reduced precision, while SMOTE-Tomek provided stability without surpassing oversampling methods in recall. Importantly, the original dataset with CEL also achieved strong recall, demonstrating the effectiveness of cost-sensitive learning even without resampling.

A key finding of this study lies in comparing the two approaches to CEL application. In Approach 1, where CEL was applied both to the original and resampled datasets, the model strongly emphasized recall. This is particularly critical for medical prediction tasks such as diabetes detection, as it reduced false negatives and increased the likelihood of correctly identifying diabetic patients. However, this improvement came with a trade-off: reduced precision, with more non-diabetic cases being incorrectly classified as diabetic.

In contrast, Approach 2 limited CEL to the original dataset while using standard CatBoost training for the resampled datasets. This approach achieved a more balanced outcome across precision, recall, and F1-score. From a practical perspective, Approach 2 is more suitable when the goal is to preserve overall classification balance, whereas Approach 1 is better aligned with high-stakes healthcare contexts where minimizing false negatives is critical, even at the cost of higher false positives.

The PR-AUC and ROC-AUC curve analyses further supported these observations. Approach 1 consistently prioritized recall, which was evident in its PR curves, while Approach 2 maintained greater stability across ROC space. These findings reinforce existing literature, which suggests that cost-sensitive learning enhances minority class detection but often compromises recall.

VII. CONCLUSION

This study assessed resampling techniques with CatBoost for diabetes prediction under class imbalance, comparing two strategies for Cost-Effective Learning (CEL). Results showed that applying CEL to both original and resampled datasets prioritize recall, making it suitable for medical settings where false negatives are costly, while restricting CEL to only the original dataset yields a more balanced overall performance, compromising recall a little. Thus, the choice of approach should depend on the application's priority: - deciding whether minimizing missed cases or maintaining predictive balance is more critical.

VIII. FUTURE WORK

Future research can expand this work in several directions. Comparing CatBoost against other ensemble methods such as XGBoost, LightGBM, and Random Forest, as well as deep learning models like LSTMs or transformers, could provide insights into generalizability across algorithms. Second, more ensemble Resampling techniques can be experimented with. For a more general understanding of these techniques, they can be applied to various other datasets that are multilabel.

REFERENCES

[1] C. BRFSS, "Kaggle," 2014. [Online]. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.

[2] N. V. B. K. W. H. L. O. &. K. W. P. Chawla, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, p. 321–357, 2002.

[3] Y. B. E. A. G. a. S. L. H. He, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, 2008.

[4] W. Y. W. a. B. H. M. H. Han, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *2005 International Conference on Intelligent Computing (ICIC)*, Hefei, China, 2005.

[5] R. C. P. a. M. C. M. G. E. A. P. A. Batista, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter,* vol. 6, pp. 20-29, 2004.

[6] A. e. a. Ramezankhani, "The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes Medical Decision Making," vol. 36, p. 137–144, 2016.

[7] A. F. N. M. a. S. R. A. Wibowo, " Refining Diabetes Diagnosis Models: The Impact of SMOTE on SVM, Logistic Regression, and Naïve Bayes for Imbalanced Datasets,," in *Journal of Electronics, Electromedical Engineering and Medical Informatics*.

[8] K. S. P. a. R. Gunasundari, "Optimizing Type II Diabetes Prediction Through Hybrid Big Data Analytics and H-SMOTE Tree Methodology," *International Journal of Computer Engineering and Smart Manufacturing (IJCESEN).*

[9] E. G. R. K. S. V. P. S. C. T. C. P. M. M. Sampath P, "Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique," in *Sci Rep*, 2024.

[10] S.-H. F.-H. J. C. M. S.-B. M.-G. M. N., "Predictive modeling of ICU healthcare-associated infections from imbalanced data. Using ensembles and a clustering-based undersampling approach," in *Applied Sciences 9*.

[11] W. K. S. L. Z. M. S. J. W. H. Yang F, "A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis," in *BMC Med Inform Decis Mak*, 2022.

[12] X. Ren, "Predictions of diabetes through machine learning models based on the health indicators dataset," in *Applied and Computational Engineering,32,216-222*, 2024.

[13] G. G. A. V. A. V. D. a. A. G. L. Prokhorenkova, "CatBoost: unbiased boosting with categorical features," *Advances in Neural Information Processing Systems (NeurIPS),* vol. 31, 2018.

[14] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, Montreal, Canada, 1995.

[15] F. W. K. S. L. e. a. Yang, "A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis," BMC Medical Informatics and Decision Making, (SMOTE-ENN results for missed abortion and diabetes datasets)," in *BMC Med Inform Decis Mak 22*, 2022.

[16] W. L. K. H. Tianyu Deng, "Identifying Key Factors that Influence Diabetes Prediction: A Meta Analysis of Two Datasets and Three Machine Learning Models," in *2nd International Conference on Machine Learning and Automation*.

[17] S. H. a. O. B. Tarid Wongvorachan, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," 2023.