Localised Ensemble Learning (LEL) - A localised approach to class imbalance

Mohamed Bader-El-Den *Member, IEEE,*, Olayemi Olabisi, James McNicholas

***Abstract*—We propose Localized Ensemble Learning (LEL), a novel framework designed to address class imbalance and data defects by focusing on localized imbalance rectification rather than relying on traditional global correction strategies. LEL leverages K-Nearest Neighbors (KNN) to categorize samples into predefined types based on specific rules, introducing a new feature, *Sample Type*, into the predictive process. Each sample type is handled independently using tailored strategies to address neighbourhood imbalance, distance imbalance, and quality imbalance, which are subsequently integrated into an ensemble model.**

The effectiveness of LEL is validated through a comprehensive evaluation against global correction strategies, such as SMOTE and Cost-Sensitive Learning (CSL), across multiple metrics including recall and precision. Statistical significance of the results is assessed using paired T-tests and Wilcoxon signed-rank tests. SHAP (SHapley Additive exPlanations) values are employed to analyze feature contributions, revealing the *Sample Type* feature as a critical determinant of model performance. Additionally, an ablation study highlights the impact of key parameters, such as the k-value in KNN providing further insights into the robustness and adaptability of the LEL framework.

Experimental results demonstrate that LEL consistently outperforms existing methods across all tested classifiers, including Random Forest, Decision Tree, XGBoost, and Naive Bayes. LEL achieves statistically significant improvements in recall and precision, underscoring its ability to handle localized forms of imbalance effectively. The findings emphasize the importance of addressing localized data defects and leveraging features like *Sample Type*, which capture complex relationships and enhance predictive accuracy.

***Index Terms*—Class Imbalance, Classification, Random Forest, Nearest Neighbour.**

I. Introduction

Class imbalance is crucial for real-world applications, particularly in high-stakes decision-making contexts. In scenarios like diagnosing life-threatening conditions or detecting financial fraud, failure to correctly predict minority class instances can lead to devastating consequences [1]. Therefore, solving the class imbalance problem is not just about improving machine learning model performance; it is about ensuring fairness and reliability in systems that directly impact human lives and businesses [1].

To address the challenges of class imbalance, several techniques have been proposed, each achieving varying degrees of success. These methods can generally be grouped into three main categories: oversampling, under sampling, and cost-sensitive learning. These techniques are typically applied at a global level, over simplifying the problem [2].

In the real world, the presence of clusters within the minority samples (small disjoints) [3] expose the limitations of oversampling, cost-sensitive approaches, and under sampling when applied at the global level. While oversampling techniques seek to enhance representation of the minority class by generating synthetic samples, they often fail to address the issue of equity. The main goal of mitigating class imbalance is that all samples become equal in the eyes of the classifier for adequate learning, unfortunately these methods can overlook the nuanced characteristics of different subgroups within the minority class when this goal is not achieved, potentially leading to synthetic samples that do not reflect the true diversity of these groups. This lack of representation can reduce the models ability to treat all subgroups fairly, ultimately hindering its ability to generalize effectively across varied contexts [4].

Cost-sensitive approaches address the miss classification costs associated with minority samples, promoting fairness in model predictions. However, when multiple clusters exist, simply increasing the weight of minority instances may not adequately resolve the issue. A uniform weighting strategy may overlook the specific needs of smaller clusters, leaving them under-represented in the learning process [5][6].

Similarly, while under sampling can mitigate the dominance of the majority class, removing samples from the majority class does not inherently ensure that the minority clusters are represented equitably, and the reduced dataset may still favour larger clusters while neglecting smaller, crucial ones [7].

Thus, when multiple clusters are present within the minority class, traditional strategies for addressing class imbalance may be insufficient. A more tailored approach is necessary to ensure that each cluster is adequately represented, preserving the unique characteristics and complexities of the minority class while fostering equitable outcomes in model predictions.

This research makes several significant contributions to the field of machine learning and class imbalance solutions. Firstly, it proposes Localised Ensemble Learning (LEL) as a flexible and reliable solution that addresses both class imbalance and other data defects by adopting a localised approach. Secondly, the research shows the performance of LEL compared to standard global methods (SMOTE and CSL), with the significance of results validated through t-Test and Wilcoxon signed rank tests. Finally, it provides deeper insights into the effectiveness of LEL's Sample Type feature, showing its significant discriminative power through SHAP (SHapley Additive exPlanations) analysis, further highlighting

Bader-El-Den and Olayemi Olabisi are with the School of Computing, University of Portsmouth, Buckingham Building, Lion Terrace, Portsmouth PO1 3HE, UK Email: Mohamed.Bader@port.ac.uk
James McNicholas is with the CriticalCareUnit, Queen Alexandra Hospital Portsmouth Hospital, and NHSTrust UK.

Manuscript received N/A; revised N/A.

the method's adaptability and precision.

The structure of this paper is organized as follows: In Section II, we explore the class imbalance problem, review state-of-the-art approaches, and discuss related works. Section **??** presents the Localised Ensemble Learning (LEL) method, providing a detailed explanation of the approach. Section IV outlines the experimental setup to support reproducibility. Section V summarizes the key findings, while Section VI delves into the implications and significance of the results. Finally, Section VII provides the concluding remarks of the paper.

II. Literature Review

A. Class Imbalance

Class Imbalance can cause substantial problems in the training and performance of machine learning models, particularly for key performance metrics such as recall and precision [8]. Models trained on imbalanced data often perform well on the majority class but struggle to accurately predict instances of the minority class, leading to poor recall (the ability to correctly identify minority class instances) meaning models may achieve high overall accuracy but fails to generalize well to the minority class, which may be the most critical category in many real-world applications [9][10][11][12][13].

Japkowicz et al. [1] provided one of the earliest systematic analyses of the class imbalance problem, exploring several resampling techniques, such as oversampling the minority class and under sampling the majority class, and demonstrated that while these methods can alleviate the problem, they are not universally effective across all domains. Building on this work, the concept of small disjoints has been introduced, referring to regions in the feature space that represent rare patterns. This has lead to a new categorisation of class imbalance approaches, namely global and localised approaches [1].

B. Global Methods

Global methods address class imbalance by treating all classes uniformly, focusing on achieving overall balance rather than adapting to the specific needs of individual subgroups. While these methods are effective at increasing minority class representation at the dataset level, they often fail to account for the nuanced characteristics of under-represented clusters, such as variations in density, positioning in feature space, and data quality. These limitations are particularly pronounced in datasets where the minority class comprises small, disjoint clusters.

A quick demonstration we designed containing 999 samples in a dataset, of which 49 belonged to the minority class, structured into two distinct clusters: cluster 1 with 37 samples and cluster 2 with 12 samples (Figure 1). After applying global oversampling techniques, 680 synthetic samples were generated, distributed as 440 for cluster 1 and 220 for cluster 2, creating a 2:1 ratio (Figure 2). This failed to respect the original 3:1 ratio, leading to inconsistent augmentation across 10 repeated trials. These results highlight how global methods disregard the unique characteristics of minority clusters, perpetuating inequities in sample distribution.

The issue of **neighbourhood imbalance**, characterized by variations in sample density within the minority class, was evident in our dataset. Clusters 1 and 2 represented distinct neighbourhoods with differing densities, yet synthetic samples were disproportionately allocated to the denser cluster. As noted by Wu [14], oversampling methods that fail to consider density disparities lead to ineffective augmentation, reducing model generalization and potentially exacerbating over fitting. Ghosh et al. [15] further demonstrated that oversampling in high-density regions without considering local neighbourhood structures increases the risk of model over fitting, particularly for larger clusters. Similarly, Chawla et al. [16] emphasized that density-aware strategies are crucial for minority class augmentation, as they help preserve the underlying data distribution and improve the classifier's ability to generalize across different regions.

In addition, **distance imbalance**, which refers to the relative positioning of samples in feature space, was observed in our experiments. Synthetic samples were often placed in sparse or irrelevant regions, particularly within cluster 2, where the original data points were less densely distributed. These misplaced samples failed to contribute meaningfully to defining decision boundaries. Liu et al. [17] highlighted that ignoring sparse regions in oversampling overlooks critical boundary cases essential for model decision-making. Additionally, Das et al. [18] found that global methods often fail to maintain feature space coherence, placing synthetic samples in regions that confuse rather than clarify the decision-making process. This misalignment between synthetic and real samples negatively impacts classifier performance and interpretability.

Finally, **quality imbalance** was evident in the amplification of noisy or low-quality samples within cluster 1. These noisy instances distorted the true decision boundary, adversely affecting the model's ability to generalize. As noted by Das et al. [18] and Lusito et al. [19], oversampling noisy data can degrade model performance by misrepresenting the true underlying data distribution. This issue is particularly problematic in small, disjoint clusters such as cluster 2, where noisy samples can disproportionately influence the synthetic data generation process. Furthermore, Rahman et al. [20] emphasized the need for quality-aware oversampling techniques that filter or exclude noisy samples to ensure the synthetic data accurately reflects the minority class characteristics.

Our findings underscore the need to move beyond global oversampling techniques. While these methods are effective at increasing overall class representation, they fail to address the intricate structures within the minority class. Addressing these limitations requires localized or adaptive approaches that respect neighbourhood characteristics, ensure proportionality across clusters, and mitigate risks associated with density, distance, and quality imbalances. Such strategies are crucial for improving model generalization and robustness, particularly in datasets with complex minority class structures.

C. Local Methods

Localized methods focus on addressing class imbalance by tailoring interventions to the structure of neighborhoods within

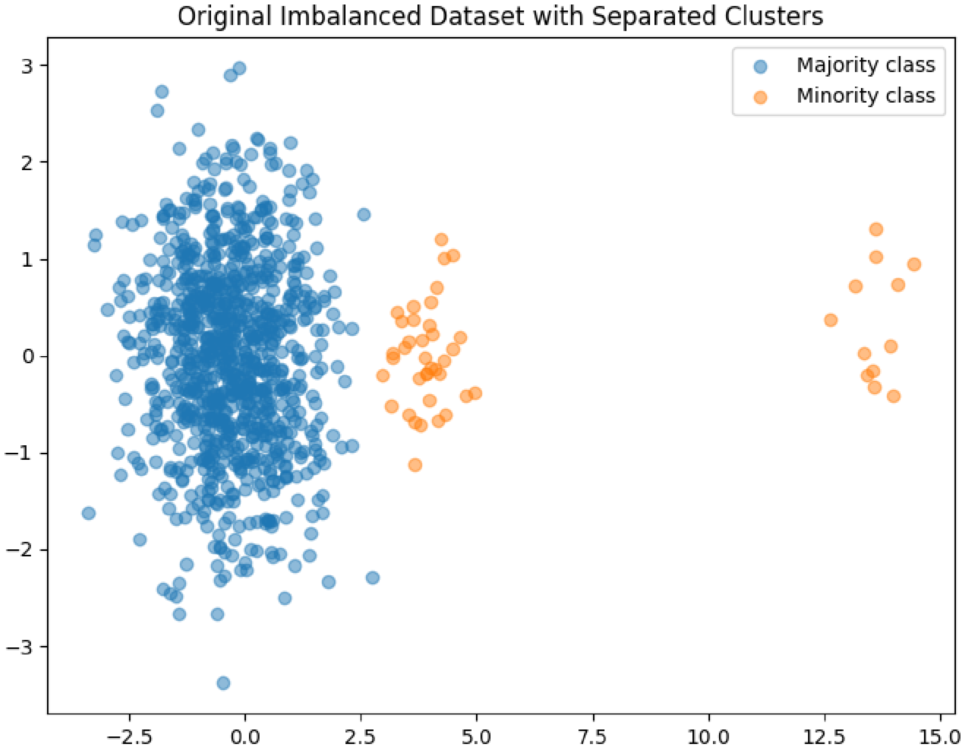

Fig. 1. Illustration of the original dataset structure highlighting the minority class distribution. The dataset contains 999 samples, with 49 belonging to the minority class. The minority class is structured into two distinct clusters: Cluster 1, comprising 37 samples, and Cluster 2, comprising 12 samples. The figure demonstrates the natural imbalance and separation within the minority class, reflecting the challenge of addressing localized variations in density and distribution through global oversampling techniques.

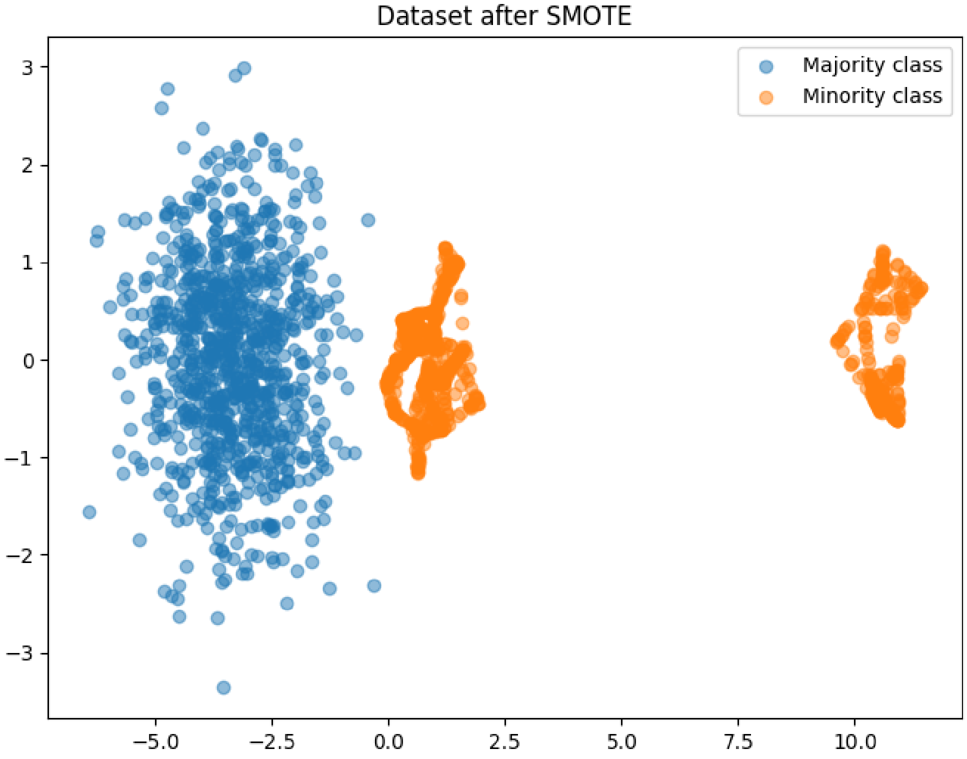

Fig. 2. Impact of global oversampling on the dataset. A total of 680 synthetic samples were generated using global methods, distributed as 440 samples for Cluster 1 and 220 samples for Cluster 2. This 2:1 ratio fails to respect the original 3:1 ratio between the clusters, resulting in disproportionate augmentation. The figure highlights how global methods amplify denser clusters while neglecting sparsely populated regions, exacerbating neighbourhood, distance, and quality imbalances within the minority class.

the data, emphasizing proximity and local density. For example, Borderline-SMOTE, introduced by [21], extends the standard SMOTE algorithm by specifically targeting ”borderline” instances in the minority class. These are samples located near the decision boundary and at a higher risk of misclassification. Synthetic samples are generated by interpolating these borderline instances with their nearest minority neighbors, under the assumption that regions near the decision boundary are most informative for improving classification performance. Studies have demonstrated that Borderline-SMOTE reduces the risk of oversampling less relevant instances compared to traditional SMOTE, proving particularly effective on smaller datasets. However, its performance can vary significantly depending on the choice of classifier and its sensitivity to decision boundaries [22][23].

Similarly, Safe-Level-SMOTE, introduced by [24], leverages the localized approach by focusing on under-represented regions of the minority class while accounting for the density and distribution of neighboring samples. This method differentiates between ”safe” and ”unsafe” areas for sample generation, ensuring that synthetic samples are primarily created in regions with sufficient minority class density. This nuanced understanding of decision boundaries enhances the likelihood of generating informative samples while avoiding noise. Experimental results show that Safe-Level-SMOTE outperforms traditional SMOTE on metrics such as AUC-ROC and F1 scores [24]. However, the complexity of calculating the safe level for each instance can introduce computational overhead and potential bias, particularly in datasets with noisy or unrepresentative local neighborhoods.

A more recent approach, Multi-label Sampling Based on Local Label Imbalance [25], introduced in 2020, applies a targeted strategy for addressing imbalance across multi-label datasets. This method selectively oversamples minority labels and undersamples majority labels based on local label distributions. By adapting sampling rates to the unique density and distribution patterns of each neighbourhood, this method demonstrates substantial improvements in Hamming loss, precision, and recall across benchmark datasets [25]. However, its reliance on computationally intensive neighborhood evaluations poses scalability challenges, particularly for large, multi-label datasets.

These approaches illustrate the strengths of localized methods in addressing critical regions of the feature space. Borderline-SMOTE and Safe-Level-SMOTE excel at leveraging proximity-based information to improve model generalization, while Multi-label Sampling Based on Local Label Imbalance underscores the importance of accounting for local distributional variations and decision boundaries. These principles align closely with concepts from [14], such as neighbourhood imbalance, which refers to variations in local sample densities, and distance imbalance, which highlights the importance of maintaining meaningful spatial relationships in feature space.

Furthermore, these methods address concerns about quality imbalance, as articulated by [14], by avoiding the amplification of noisy or low-quality samples that could distort decision boundaries. For instance, Safe-Level-SMOTE explicitly aims to avoid generating synthetic samples in unsafe regions, mitigating the risk of over fitting or introducing noise.

As corroborated by other studies, localized methods hold significant potential for improving classifier performance in imbalanced datasets. [26] emphasize the necessity of understanding local class distribution patterns, while [27] and [28] highlight the role of localized oversampling in enhancing decision boundary precision. However, challenges such as computational complexity and sensitivity to noisy neighbourhoods remain. Despite these limitations, localized methods offer a promising avenue for addressing nuanced forms of imbalance, paving the way for more robust and adaptive models.

III. Localised Ensemble Learning (LEL)

This section introduces the **Localized Ensemble Learning (LEL)** approach, motivated by the observation that conventional class imbalance mitigation techniques operate predominantly at a global level [29], [30], [31]. These global approaches, while effective in addressing overall class imbalance, often fail to account for the nuanced distributions and localized characteristics of minority classes [25]. Inspired by this limitation, researchers have shifted their focus toward localized methods, which prioritize identifying and preserving the unique patterns of minority instances within their local neighborhoods [25]. LEL incorporates this philosophy into a three-phase process:

1) **Phase 1:** Identification of sample types (Safe, Borderline, Rare) within the dataset (Figure 3).
2) **Phase 2:** Customized treatment of each sample type using targeted techniques.
3) **Phase 3:** Construction of an ensemble that emphasizes minority class learning through algorithm-level oversampling.

A. Identifying Sample Types

Phase 1 involves classifying instances in the dataset $\mathcal{X}$ based on their local neighborhood structure. For each instance $x_i \in \mathcal{X}$, the minority neighborhood density $n_m(x_i)$ is determined using k-nearest neighbors (k-NN) with a chosen distance metric (e.g., Euclidean distance). This is formalized as:

$$n_m(x_i) = \sum_{x_j \in \mathcal{N}_k(x_i)} I(x_j \in \text{Minority}), \tag{1}$$

where $\mathcal{N}_k(x_i)$ represents the k-nearest neighbors of x_i, and $I()$ is an indicator function that equals 1 if x_j belongs to the minority class and 0 otherwise. The minority-to-majority ratio $r(x_i)$ is then calculated as:

$$r(x_i) = \frac{n_m(x_i)}{n_M(x_i) + 1}, \tag{2}$$

where $n_M(x_i)$ denotes the count of majority instances in the neighborhood. Based on threshold values t_s, t_b, and t_r (inspired by [25]), instances are categorized as follows:

- **Rare Instances:** Minority-to-majority ratio $r(x_i)$ falls below t_r, indicating the instance resides in a region dominated by majority samples, making it vulnerable to misclassification.
- **Safe Instances:** Ratio $r(x_i)$ meets or exceeds t_s, indicating the instance is within a dense cluster of minority samples, offering strong support for correct classification.
- **Borderline Instances:** Ratio $r(x_i)$ lies between t_b and t_s, indicating the instance resides near a decision boundary with comparable representation from both classes.

The sample type classification can be expressed as:

$$\text{Sample Type } x_i = \begin{cases} \text{Safe} & \text{if } n_m(x_i) \geq t_s, \\ \text{Borderline} & \text{if } t_b \leq n_m(x_i) < t_s, \\ \text{Rare} & \text{if } n_m(x_i) < t_r. \end{cases} \quad (3)$$

B. Customized Treatment of Sample Types

Once categorized, each sample type is addressed using tailored techniques designed to mitigate specific challenges as an example, below is a possible configuration, although ideally experiments needs to be executed to inform the chosen configuration as seen in Section III-D :

- **Rare Instances:** Can be treated using *SMOTE* [32], which generates synthetic samples to enhance the representation of these sparse regions. SMOTE ensures that the minority class is better represented without over fitting.
- **Safe Instances:** Can be addressed through *random undersampling*, which removes redundant majority instances. This reduces model over fitting to easy-to-classify examples, sharpening the decision boundary [33].
- **Borderline Instances:** Could be managed with *Tomek Links* [34], which clean ambiguous instances near decision boundaries. This process reduces class overlap and sharpens the boundary for better minority class representation.

By applying these techniques selectively, LEL leverages the strengths of each method while minimizing their limitations, improving overall model performance.

C. Ensemble Construction with Algorithm-Level Oversampling

In Phase 3, LEL builds an ensemble of which its components are the classifiers with unique class mitigation strategies addressing each sample type.

Localized Ensemble Learning (LEL) represents a significant advancement over traditional global methods by addressing class imbalance through the lens of localized imbalance rectification. By explicitly targeting **neighbourhood imbalance**, LEL tailors interventions to specific sample types within minority class regions, ensuring that augmentation directed where necessary. This approach aligns with the principles of proximity-based interventions discussed by Wu [14], allowing for a more nuanced handling of variations in local sample densities.

In addition, LEL addresses **distance imbalance** by refining decision boundaries through localized learning strategies. This ensures that the placement of synthetic samples and classifier adjustments occur in meaningful regions of the feature space, preserving spatial relationships that are critical for robust decision-making. By focusing on these meaningful regions, LEL reduces the risk of overfitting and enhances generalization, particularly in datasets characterized by sparse or complex minority class distributions.

LEL also incorporates mechanisms to mitigate **quality imbalance**, prioritizing high-quality samples and avoiding the amplification of noisy or low-quality instances that could distort decision boundaries.

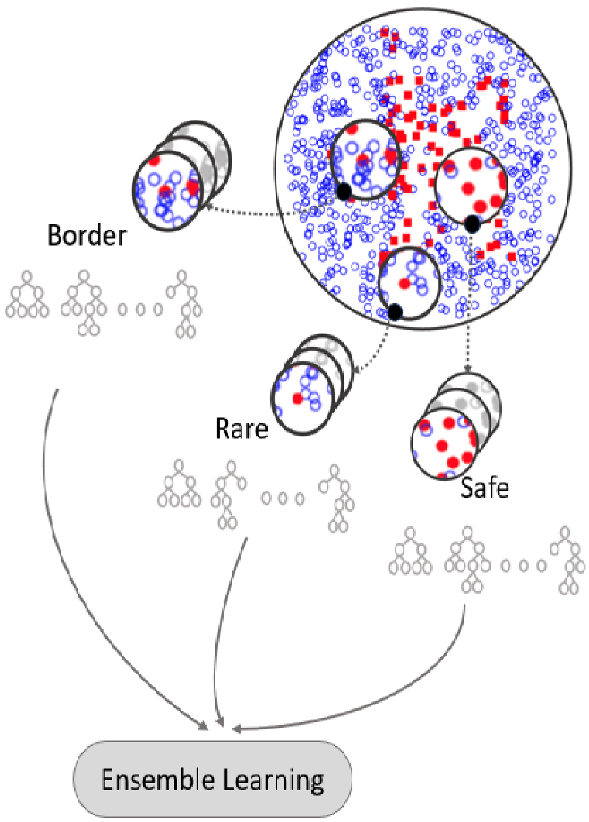

Fig. 3. Visualization of the Localized Ensemble Learning (LEL) framework's. The figure illustrates the categorization of instances in the dataset into three distinct sample types â Safe, Border, and Rare â based on their local neighbourhood structure and minority-to-majority ratio ($r(x_i)$). Safe instances are located within dense minority or majority clusters, offering robust support for correct classification. Border instances are positioned near decision boundaries, where class overlap is significant, while Rare instances reside in regions dominated by opposite class samples, making them highly susceptible to misclassification. This process forms the foundation for tailored treatments in LEL, ensuring each sample type is addressed with methods specifically designed to enhance minority class representation which is then combined into an ensemble.

D. Ablation Study

Localised Ensemble Learning (LEL) is made up of various phases as outlined above, in this section, we investigate the impact of each parameter to the overall framework, these include the K-value, sample types generated, ensemble bias ratios, and class imbalance mitigation techniques. A Naive Bayes classifier was used as the base model on the P3 dataset to demonstrate the effects of these configurations on model performance.

K**-Value**: Literature consistently shows that the K-value is a key parameter influencing prediction accuracy. Studies highlight that systematic tuning of K, through methods such as grid search or cross-validation, rather than arbitrary selection or educated guesses, leads to notable improvements in model performance [35], [36]. We test$K$-values of 5, 10, 15 and 20.

Sample Types Definition: The specified sample types can also impact the performance and effectiveness of LEL. Since LEL leverages a localized approach, increasing sample type categories can enhance the modelâs ability to capture distinct clusters in the data. However, prior research suggests that a threshold exists, beyond which additional sample types yield diminishing returns in performance and add computational complexity [37], [38]. This study doesn't explore this but uses a fixed ratio where safe is defined as samples surrounded by at least 80% of its own class, Rare is defined as instances surrounded by less than 20% of its own class, borderline is assigned when the sample is surrounded with between 40% to 60% of its own class and lastly a group named Mixed for everything else.

Class Imbalance Mitigation Strategies: To address class imbalance, a range of techniques were randomly selected from a comprehensive list. Some of which include the Synthetic Minority Over-sampling Technique (SMOTE) [39], Adaptive Synthetic Sampling (ADASYN) [40], NearMiss, Random OverSampler, Random Undersampler, SMOTETOMEK, SMOTEENN. These methods include both oversampling techniques, which generate synthetic samples for the minority class, and under-sampling techniques, which reduce the majority class size to achieve balanced class distributions. Applying these strategies enables the model to handle imbalanced data more effectively, enhancing both predictive accuracy and reliability for minority classes.

Ensemble Bias Ratio: The ensemble bias ratio determines the weight or bias assigned to each sample type within the LEL model. We hypothesize that adjusting this ratio can significantly affect classifier performance. Effective calibration of the ensemble bias ratio is crucial for balancing contributions from different sample types, thereby improving prediction accuracy and model robustness [41], [42].

E. Ablation Study Results

The experiments conducted evaluated the different parameters of the LEL approach namely, class mitigation strategies, k-values and sample type ratios. The key metrics evaluated were **Recall** and **Precision**, with the goal of maximizing both across different experimental setups, the results are displayed in Table V.

a) Experiment 1: The configuration was **NearMiss** strategy applied to Safe and Rare sample types while **Random Oversampling** applied to borderline. $k = 5$ yielded the highest recall (**0.4657**) and precision (**0.1306**). As k increased, a slight decline was observed in both metrics, with the lowest performance recorded at $k = 20$ (Recall $= 0.4561$, Precision $= 0.1276$).

b) Experiment 2: The configuration was **SMOTE** for Safe, **Random Oversample** for Borderline, and **TomekLinks** for Rare samples. $k = 10$ achieved the best overall results with a recall of **0.8739** and precision of **0.8670**. High performance was observed across all k-values.

c) Experiment 3: The configuration is **SMOTEENN** for Safe, **CondensedNearestNeighbour** for Borderline, and **Random Oversample** for Rare samples. $k = 5$ resulted in a recall of **0.8719** for and precision of **0.8549** which is the best. Performance decreased as k increased, with $k = 20$ yielding recall and precision values of 0.8582 and 0.6860.

d) Experiment 4: The configuration for experiment 4 include **CondensedNearestNeighbour** for Safe, **SMOTETomek** for Borderline, and **ADASYN** for Rare samples. The best recall (**0.8720**) and precision (**0.8603**) was achieved with $k = 5$. As k increased to 20, performance declined slightly (Recall $= 0.8520$, Precision $= 0.6750$).

F. Ablation Study Analysis

The full table outlining the results of the 4 experiments carried out are seen in Table V. Below are interesting insights.

- **SMOTEENN** and **SMOTETomek** demonstrated robustness across various configurations, particularly in Experiments 2 and 3.

  SMOTEENN and SMOTETomek are hybrid oversampling and undersampling methods that combine the strengths of SMOTE with noise-cleaning techniques. SMOTE generates synthetic samples for the minority class, addressing under-representation, while ENN (Edited Nearest Neighbor) and Tomek links remove noisy samples or overlapping majority-class samples, improving class separability. Their robustness in Experiments 2 and 3 could be attributed to their ability to simultaneously enhance minority-class representation while mitigating the influence of noisy or ambiguous samples. This balance is particularly effective in datasets with complex decision boundaries or overlapping classes, where traditional oversampling methods may inadvertently amplify noise. These methods help refine the decision space, leading to better generalization and improved model performance.
- **The ensemble ratio, introduced as a means to add extra bias, appears to have no influence on the outcome of recall and precision.**

  The lack of influence of the ensemble ratio on recall and precision suggests that the modelâs performance is primarily driven by the underlying data distributions and sampling strategies rather than the specific weighting of classifiers within the ensemble. This could indicate that the ensemble methods are robust enough to achieve a consistent level of performance regardless of the relative contributions of individual classifiers. It is also possible that the ensemble ratio does not significantly alter the diversity of decision boundaries created by the base learners, which might explain the negligible effect on the outcome metrics.

- **Lower k-values yield better results.**
  The observed trend of lower k-values yielding better results is particularly relevant to the generation of the "Sample Type" feature. The "Sample Type" feature encapsulates underlying patterns and relationships between features. Using a smaller k-value results in a more nuanced representation of minority class data, effectively capturing its heterogeneity and avoiding oversimplified interpolations. Conversely, higher k-values may cause the generated "Sample Type" feature to incorporate less relevant or overly generalized patterns, diluting its discriminatory power and reducing its overall utility. By maintaining tighter local relationships, lower k-values contribute to a "Sample Type" feature that better reflects meaningful variations in the data, improving model interpretability and performance.

IV. Experimental Setup

In this section, we explained the performance and comparison of LEL to traditional class imbalance mitigation strategies as well as base classifiers, using a scientific methodology to ensure reproducible and valid results.

A. Datasets

To be able to better evaluate the performance of LEL, a set of existing and new datasets are considered in this study. The set consists of 8 datasets; 3 derived (P3,P6,P12) from physio net [43], and 5 classic class imbalance dataset from the KEEL repository [44]. Summary of the datasets are available in table IV-A. The following subsections go through the details of each of these datasets.

TABLE I
Summary of Dataset. Imbalance Ratio (IR)

Dataset	Attributes	Number of Records	IR
P3	100	280292	6.0
P6	100	280292	5.0
P12	100	262700	4.0
Haberman	3	306	2.7
BUPA	6	345	2.37
POKER	10	2076	82
PIMA	8	768	1.86
CONS	43	40657	20.56

1) P3, P6, P12 Datasets: The Physio-net 2019 data set [43] was used in this study to generate 3 datasets for the purpose of this study. The increasing prevalence of sepsis in clinical settings has raised significant concerns regarding early detection and intervention [45]. The low incidence rate of sepsis among hospitalized patients creates a significant class imbalance [45]. According to the Sepsis-3 guidelines, the incidence of sepsis ranges from 1% to 3% of hospitalized patients, depending on the population studied. The clinical consequences of failing to accurately predict sepsis are dire. Mortality rates for septic patients can exceed 30% if not promptly identified and treated [46].

The physio net dataset [43] encompasses 40,336 patients, of which 2,932 patients experienced sepsis during their hospitalisation. Each patient included in the dataset has been annotated at hourly timestamps with either a positive or negative label. As defined by [43], a positive label indicates a clinical suspicion of infection, supported by medical events such as the administration of antibiotics and lab culture testing within a specified time frame, or an official diagnosis of sepsis based on a two-point increase in the Sequential Organ Failure Assessment (SOFA) score.

However, it is crucial to acknowledge that the dataset is afflicted by a considerable number of missing values, with all 1,552,210 records having at least one variable with missing data. To offer further insights into the dataset, Table II presents detailed information about the features.

To enhance the data utility for prediction purposes, we apply the sliding window technique which was originally developed to facilitate the processing of large inputs and sequential data [47]. We created parameters for the sliding window algorithm, which include Window Size, Gap Size, Prediction Size, and Step Size. The Prediction Size refers to the time in advance for which a specific event will be predicted. The reading window defines the time span for which data will be aggregated to make the prediction. The Step Size determines the starting point for the next reading window, while the Gap Size determines the time gap between the reading window and the prediction window. The output for the sliding window include maximum value, minimum value, first value, last value, average value, mean value and standard deviation. An illustration of these parameters can be seen in Figure 4.

In this study, 3 distinct data sets, namely P3, P6 and P12, were generated using the sliding window algorithm. Data set P3 was generated using a Reading window of 3, Gap Size of 0, Step Size of 0, and Prediction Size of 3. On the other hand, data set P6 was generated using a Reading window of 3, Gap Size of 0, Step Size of 0, and Prediction Size of 6 and P12 was generated using a Reading window of 3, Gap Size of 0, Step Size of 0, and Prediction Size of 12.

2) KEEL Repository Datasets: The KEEL (Knowledge Extraction based on Evolutionary Learning) repository is a comprehensive resource specifically designed for researchers and practitioners working on data mining and machine learning, with a particular focus on class imbalance problems [44].

Researchers can utilize KEEL to develop, test, and compare different approaches for handling class imbalance, enabling more robust and generalizable solutions across a wide range of real-world applications, including medical diagnosis, fraud detection, and risk prediction [44].

In this study, we also used the Haberman dataset, created by Dr. William H. Wolberg, includes 306 instances of breast cancer patients who underwent surgery at the University of Chicago's Billings Hospital between 1958 and 1970, featuring attributes such as age at surgery, year of operation, positive auxiliary nodes, and survival status (1 = survived over 5 years; 2 = deceased within 5 years). The BUPA dataset, which explores the relationship between various health attributes and liver disorders, including liver enzyme levels and daily

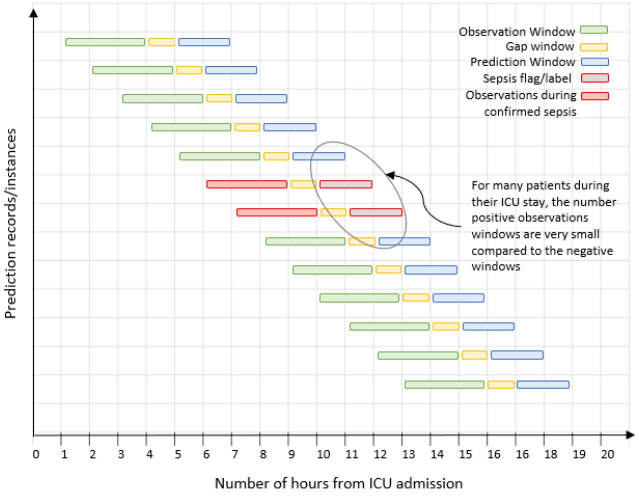

Fig. 4. Visualization of the sliding window parameters used for dataset generation. The sliding window algorithm processes sequential data by defining specific parameters: Window Size, Gap Size, Prediction Size, and Step Size. The **Reading Window** aggregates data over a specified time span to make predictions. The **Prediction Size** indicates the time in advance for which the event (e.g., sepsis) is predicted. The **Gap Size** defines the interval between the Reading Window and the Prediction Window, while the **Step Size** determines the starting point for the next Reading Window. The algorithm outputs statistical features such as maximum, minimum, first, last, mean, average, and standard deviation values. This algorithm was used to generate the P3, P6, and P12 datasets for sepsis prediction, with varying Prediction Sizes of 3, 6, and 12 hours, respectively.

alcohol consumption, with a binary target variable indicating the presence (1) or absence (2) of liver disorders. The Poker dataset, which contains data on poker hands, consisting of ten features related to card suits and ranks, categorized into ten classes representing different poker hands. The PIMA dataset, which focuses on female Pima Indian patients, examining health factors related to diabetes, with attributes like pregnancies and glucose levels, and a binary target indicating diabetes presence (1) or absence (0). Finally, the CONS dataset captures consumer decision-making, featuring attributes such as age, gender, and purchasing history, with a binary class label indicating a positive (1) or negative (0) purchase decision [44].

B. Data Pre-Processing

For datasets to be effective in producing accurate models, key prepossessing techniques must be applied. These include handling missing values, feature reduction, and normalization, among others. Each of these techniques is crucial to enhancing the quality and utility of the data, and they are explained in detail below.

1) Feature Reduction: Feature reduction is an essential step in data preprocessing that helps eliminate irrelevant or redundant features which leads to improved model performance and reduced computational complexity. Various methods can be used for feature reduction, including Principal Component Analysis (PCA), L1 Regularization (Lasso), Recursive Feature Elimination (RFE) , Information Gain among others [48].

In this research, Information Gain (specifically using the Gini index) has been adopted as the feature reduction method

TABLE II
ORIGINAL FEATURES WITH CORRESPONDING STATISTICS

	Features	Miss	Plausibility		Mean	Med	STD
		%	L	H			
Vital Signs	HR	9.9	10	300	84.6	83.5	17.3
	O2SAT	13.1	60	100	97.2	98	2.7
	TEMP	66.2	32	42.2	37	37	0.8
	SBP	14.6	40	280	123.8	121	23.2
	MAP	12.5	0	300	82.4	80	16.2
	DBP	31.3	20	130	63.7	62	13.6
	RESP	15.4	5	60	18.8	18	5
	ETC02	96.3	0	150	33	33	8
Laboratory Values	Base Excess	94.6	-20	20	-0.7	0	4.2
	HCO3	95.8	0	50	24.1	24	4.4
	FI02	91.7	0	1	0.5	0.5	0.2
	PH	93.1	6	8	7.4	7.4	0.1
	PACO2	94.4	0	200	41	40	9.3
	SAO2	96.5	0	100	92.7	97	10.9
	AST	98.4	0	400	64.5	35	73
	BUN	93.1	0	500	23.9	17	20
	ALKALINE_P	98.4	0	250	82.8	71	43
	CALCIUM	94.1	0	20	7.6	8.3	2.4
	CHLORIDE	95.5	75	145	105.8	106	5.8
	CREATININE	93.9	0	10	1.4	0.9	1.4
	BILIRUBIN(D)	99.8	0	50	1.8	0.4	3.7
	GLUCOSE	82.9	0	1000	136.9	127	51.3
	LACTATE	97.3	0	100	2.6	1.8	2.5
	MAGNESIUM	93.7	0	10	2.1	2	0.4
	PHOSPHATE	96	0	12	3.5	3.3	1.4
	POTASSIUM	90.7	1	10	4.1	4.1	0.6
	BILIRUBIN(T)	98.5	0	50	2.1	0.9	4.3
	TROPONIN I	99	0	200	8	0.3	22.7
	HCT	91.1	10	70	30.8	30.3	5.5
	HGB	92.6	2	22	10.4	10.3	2
	PTT	97.1	0	250	41.2	32.4	26.2
	WBC	93.6	0	50	11.2	10.3	5.4
	FIBRINOGEN	99.3	0	800	280.2	248	137.5
	PLATELETS	94.1	5	1500	196	181	103
Demographics	AGE	0	0	150	62	64	16.4
	GENDER	0	0	1	0.6	1	0.5
	UNIT1	39.4	0	1	0.5		0.5
	UNIT2	39.4	0	1	0.5	1	0.5
	Duration	0			-56.1	-6	162.3
	HCULOS	0	1		27	21	29

due to its ability to evaluate the contribution of each feature based on how well it differentiates between classes, using impurity reduction in decision trees. This method is particularly useful for healthcare datasets, which often contain numerous features, some of which may be redundant or irrelevant. Healthcare data can also be complex, with features that vary significantly in importance depending on the condition or outcome being studied [48]. By focusing on features that maximize predictive power while minimizing noise, this method can even aid in discovering key insights into the factors that influence patient outcomes [49].

2) Handling Missing Values: Given the health-related focus of the datasets under examination and the prevalence of missing values as seen in Figure 5, various techniques were investigated as the criticality of the choice of data imputation method cannot be overstated [50]. Thus, the selection of imputation techniques aligns with the specific objectives of the research task. In this regard, the Forward Fill and Backward Fill methods were adopted.

Forward Fill (denoted as FFill) and Backward Fill (denoted

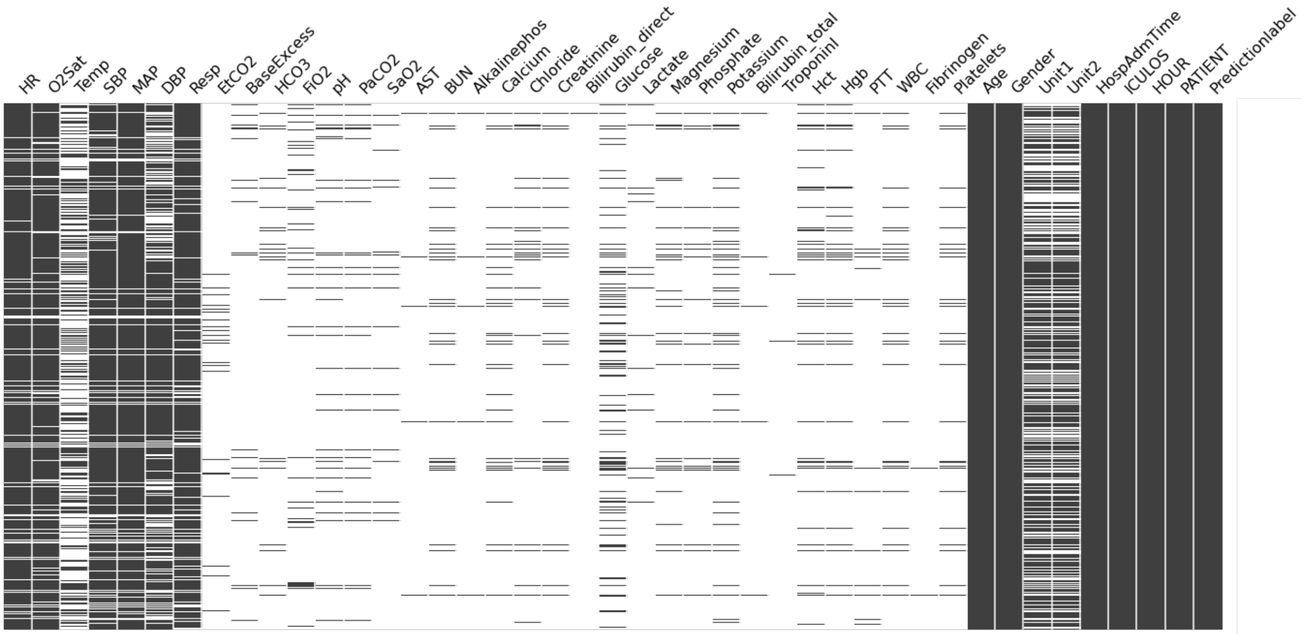

Fig. 5. Visualization of the features provided in the original PhysioNet 2019 dataset for sepsis prediction, highlighting the proportion of missing values for each feature. The dataset includes clinical measurements annotated at hourly intervals, with significant variation in data completeness across features. This visualization underscores the challenge posed by missing values, which affects all 1,552,210 records in the dataset. Such missingness necessitates preprocessing techniques to ensure data utility and reliable sepsis prediction.

as BFill) are widely recognized techniques for addressing missing values, particularly within the context of time-series or sequential data analysis. FFill works by propagating the most recent non-missing observation forward along a specified axis. This imputation strategy presupposes that the most recent data point represents a credible approximation for the absent value. In contrast, BFill involves substituting missing values with the nearest past non-missing observation along the same axis. Both methods are frequently applied in the analysis of time-series data where the datasets may exhibit gradual temporal trends or patterns. Both Forward Fill and Backward Fill serve as indispensable tools for managing missing data, allowing the preserving of dataset integrity for subsequent analytical and modelling endeavours [50].

3) Normalisation: Normalization is a fundamental data prepossessing technique that plays a pivotal role in ensuring the effectiveness of various analytical and modeling methods. The goal of normalisation is to transform numerical features into a consistent and standardized range, typically [0, 1] or [-1, 1]. By doing so, it eliminates the influence of feature magnitude, allowing models to give equal weight to each feature during training. This process enhances model stability, convergence speed, and the interpret-ability of model parameters, ultimately resulting in more reliable predictions [51].

In this paper, we opted to apply min-max scaling for normalisation. Min-Max scaling is one of the most widely used normalisation techniques, linearly scales feature values to fit within a specified range, typically [0, 1]. The formula for Min-Max scaling is as follows [52]:

$$X_{\text{normalised}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{4}$$

The simplicity and effectiveness of this approach make it ideal for this work [53].

C. Algorithm

In this study, we utilized Random Forest, Decision Tree, XGBoost, and Naive Bayes as base classifiers. Each algorithm provides a unique approach to classification, allowing a comprehensive evaluation of the effectiveness of LEL approach.

Random Forest: Random Forest is an ensemble classification algorithm that combines multiple decision trees to improve prediction accuracy and reduce over fitting [54]. Each tree T_b in the forest is constructed from a bootstrapped sample of the data and trained using a subset of features at each node. This feature randomness reduces correlation between trees and enhances generalization [55].

The prediction for a new instance is determined by aggregating predictions from all B trees:

$$\hat{y} = \text{mode}\left(\hat{y}_1(x), \hat{y}_2(x), \ldots, \hat{y}_B(x)\right),$$

where $\hat{y}_b(x)$ is the predicted class from the b-th tree.

Decision Tree: Decision Tree classify data by recursively partitioning the feature space based on criteria that maximize node purity [56]. At each node, a feature is selected to split the data based on metrics such as:

- Gini Impurity:

$$G(p) = 1 - \sum_{k=1}^{K} p_k^2,$$

where p_k is the probability of class k within the node.

- Entropy:

$$H(p) = -\sum_{k=1}^{K} p_k \log_2(p_k),$$

where p_k is the class probability at the node.

For classification, Information Gain (IG) is used to evaluate the quality of each split:

$$IG = H(p) - \sum_{j=1}^{m} \frac{|S_j|}{|S|} H(p_j),$$

where, $H(p)$ is the entropy of the parent node, S_j represents each subset after the split, and $|S_j|$ is the size of subset j.

The predicted class for a new instance is determined by the majority class in the leaf node it reaches.

Naive Bayes: Naive Bayes is a probabilistic classifier based on Bayes' Theorem, assuming independence between features [57]. The posterior probability of class C given feature vector $X = (x_1, x_2, \ldots, x_n)$ is calculated as:

$$P(C|X) \propto P(C) \prod_{i=1}^{n} P(x_i|C).$$

Types of Naive Bayes classifiers include:

1. Gaussian Naive Bayes (for continuous data), which assumes a normal distribution for features:

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} \exp\left(-\frac{(x_i - \mu_C)^2}{2\sigma_C^2}\right),$$

where μ_C and σ_C^2 are the mean and variance for feature x_i in class C.

2. Multinomial Naive Bayes (for count data), where $P(x_i|C)$ is calculated as:

$$P(x_i|C) = \frac{\text{count}(x_i, C) + \alpha}{\sum_j \text{count}(x_j, C) + \alpha \cdot V},$$

with V as vocabulary size and α as a smoothing parameter.

Classification is performed by selecting the class with the highest posterior probability:

$$\hat{C} = \arg\max_{C_k} P(C_k) \prod_{i=1}^{n} P(x_i|C_k).$$

XGBoost: XGBoost (eXtreme Gradient Boosting) is a scalable, optimized implementation of gradient boosting, widely used for classification due to its efficiency and accuracy [58]. The objective function for XGBoost consists of a loss function L and a regularization term Ω:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k),$$

where f_k represents each decision tree, and $\Omega(f_k)$ controls the model complexity.

Using a second-order Taylor expansion, XGBoost approximates the objective function for efficient optimization:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[L(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t),$$

where g_i and h_i are first and second derivatives of L with respect to the prediction $\hat{y}_i^{(t-1)}$.

Each tree is built by selecting splits that maximize the gain:

$$\text{Gain} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma,$$

where G_L, H_L, G_R, and H_R are sums of gradients and Hessians for the left and right child nodes, respectively.

D. Performance Metrics

In classification tasks, the evaluation of model performance commonly relies on accuracy, which quantifies the proportion of correctly predicted instances. However, accuracy is unsuitable for scenarios involving class imbalance, as it tends to exhibit a bias towards the majority class and may yield a high accuracy percentage even if all the minority examples are misclassified [59]. As the focus of interest lies on the minority class, accuracy would not be a suitable choice of metric in this study.

Unlike Accuracy, precision and recall are crucial metrics because they allow better insight into the performance of the classifier on minority classes. Precision measures the proportion of true positive predictions out of all predicted positives [59]. In imbalanced datasets, high precision ensures that the minority class predictions are relevant and not overwhelmed by false positives. This is particularly useful in applications where false alarms are costly, such as fraud detection or medical diagnosis. Recall (also called sensitivity) measures the proportion of true positive predictions out of all actual positives [59]. High recall ensures that most of the actual minority class instances are detected. This is critical in situations where missing a minority class instance (e.g., a rare disease) is more harmful than making false positive predictions. By focusing on precision and recall, we can ensure a better assessment of model's ability to handle imbalanced data, ensuring that minority class instances are not ignored or misclassified [59].

E. LEL Parameters

1) K- Values Sample Type: In order to enable the custom treatments of sample types as described in Section III,. KNN is applied to the dataset to identify the different types of sample, this sample type is then added to the dataset and named "Sample Type". This label is both used to treat each type ("safe","borderline" and "rare") uniquely but also is used for prediction. In this paper, K = 5 is used and was determined based on the ablation study conclusions in III-F. The "Sample Type" label is then generated based on the rules below:

Let:

- m: the count of majority class samples in the neighborhood.
- n: the count of minority class samples in the neighborhood.

The classification can be determined by the following conditions:

A sample is classified as safe if the neighborhood is dominated by samples from the same class, either all majority or all minority.

- If $n = 5$ and $m = 0$: All samples are from the minority class, thus the sample is "Safe" for the minority class.

$$\text{Safe} \quad \text{if} \quad (n = 5 \wedge m = 0)$$

- If $m = 5$ and $n = 0$: All samples are from the majority class, thus the sample is "Safe" for the majority class.

$$\text{Safe} \quad \text{if} \quad (m = 5 \wedge n = 0)$$

A sample is classified as rare when it is significantly outnumbered by the other class, indicating a rare occurrence of the minority class in a mostly majority-class neighbourhood or vice versa.

- If $m = 4$ and $n = 1$: The sample is minority, with only one minority class sample surrounded by four majority class samples.

$$\text{Rare} \quad \text{if} \quad (m = 4 \wedge n = 1)$$

- If $m = 1$ and $n = 4$: The sample is majority, with only one majority class sample surrounded by four minority class samples.

$$\text{Rare} \quad \text{if} \quad (m = 1 \wedge n = 4)$$

A sample is classified as borderline when the neighborhood has a relatively balanced distribution of majority and minority classes, indicating that the sample lies near the boundary between the two classes.

- If $m = 3$ and $n = 2$: The sample is minority and lies in a neighborhood where majority samples are slightly more common.

$$\text{Borderline} \quad \text{if} \quad (m = 3 \wedge n = 2)$$

- If $m = 2$ and $n = 3$: The sample is majority and lies in a neighborhood where minority samples are slightly more common.

$$\text{Borderline} \quad \text{if} \quad (m = 2 \wedge n = 3)$$

Once the sample type is generated, each type is treated with ideal class imitigation stratgies outlined below.

2) Class Mitigation Strategy: The ablation study presented in Table V provides critical insights into the impact of various class mitigation strategies on different sample types. The results clearly indicate that a tailored approach is necessary to optimize recall and precision, with rare samples benefiting significantly from oversampling, while safe and borderline samples can be effectively managed using undersampling methods. These findings will inform the parameter selection for Localised Ensemble Learning (LEL) in the main experiments.

The ablation study reveals substantial variations in model performance across different strategies. In Experiment 1, where NearMiss was applied uniformly to all sample types, recall and precision remained low, these results demonstrate that undersampling alone is insufficient to handle all sample complexities, particularly for rare samples, which remain under-represented in the decision boundary regions.

In contrast, Experiments 2, 3, and 4, which employed oversampling techniques such as SMOTE, SMOTEENN, and ADASYN, showed a remarkable increase in both metrics. For example in experiment 3 where SMOTEENN yielded 0.8719 recall and 0.8549 precision, while Experiment 4 using ADASYN produced 0.8720 recall and 0.8603 precision. These findings highlight that rare and samples significantly influence false negatives and thus require oversampling to improve classifier generalisation.

Safe and borderline samples, however, displayed minimal performance fluctuations across different strategies, suggesting that oversampling offers limited benefit for these regions. These samples, typically located near or within stable class boundaries, can be adequately handled by reducing the size of the majority class using methods such as NearMiss or Condensed Nearest Neighbour (CNN).

To understand these results, it is essential to examine the underlying mechanisms of the applied strategies. SMOTE addresses class imbalance by generating synthetic samples for the minority class through linear interpolation between existing samples and their nearest neighbours. This method increases the density of minority samples in sparse regions, particularly benefiting rare samples. By reinforcing the decision boundary, SMOTE reduces the likelihood of misclassifications, thereby improving recall and reducing false negatives. However, SMOTE may introduce noise if safe and borderline samples are over-synthesised, which is why it is less effective for these sample types.

NearMiss selectively reduces the majority class by retaining only those samples that are closest to the minority class, emphasising boundary regions. This strategy is particularly suited for safe and borderline samples, where oversampling is unnecessary. By focusing on the most informative points near decision boundaries, NearMiss creates a compact and balanced dataset, preventing overfitting and improving precision without generating synthetic data.

Based on the ablation study findings, the following approach will be adopted in the main experiments to optimise LEL parameters: Safe and borderline samples will be handled using NearMiss to undersample the majority class. This will ensure that the dataset remains balanced and compact, with only the most informative samples retained near the decision boundary. Rare and outlier samples will be managed using SMOTE, which generates synthetic samples to amplify their representation in the dataset. This approach will preserve and strengthen decision boundaries, improving the model's ability to generalise across underrepresented regions.

In conclusion, the ablation study provides a comprehensive justification for the use of differentiated class mitigation strategies in Localised Ensemble Learning. Rare and outlier samples require targeted oversampling to reduce false negatives, while safe and borderline samples benefit from undersampling to prevent overfitting. This tailored approach ensures that the model can generalise effectively across all sample types, ultimately improving its overall performance.

3) Ensemble Ratio: The results in Table V indicate that variations in the ensemble ratio of sample types (safe, borderline, and rare) have little to no influence on recall and precision. This is evident across all experiments and strategies, where the recall and precision remain relatively stable regardless of changes in the number of each sample type present in the ensemble.

For instance, when the ratios of safe, borderline, and rare samples were altered in different configurations (e.g., 6:1:1, 1:6:1, 1:1:6, 1:1:1), there was no notable impact on the modelâs performance metrics. In Experiment 1, the recall consistently hovered around 0.4561â0.4657, and precision remained at 0.1276â0.1306, regardless of how the ensemble ratio was adjusted. Similarly, in Experiment 2, where SMOTE and Tomek Links were applied, recall and precision stayed at 0.8736â0.8739 and 0.8661â0.8670, respectively, across varying ratios.

This trend is further observed in Experiments 3 and 4, where SMOTEENN and ADASYN were used. Despite altering the ratio of safe, borderline, and rare samples, recall and precision remained consistent at approximately 0.8719â0.8720 and 0.8549â0.8603, respectively. This suggests that the ensemble ratio does not play a significant role in determining the final performance outcomes once appropriate class mitigation strategies are in place.

These findings indicate that the performance improvements are primarily driven by the choice of class imbalance mitigation techniques rather than the sample type ratios within the ensemble. Effective strategies such as SMOTE for rare samples and NearMiss for safe and borderline samples ensure that critical points near decision boundaries are appropriately represented, making the exact distribution of these sample types less critical. This highlights the robustness of the hybrid strategy proposed for Localised Ensemble Learning, where handling each sample type with a suitable mitigation method can yield consistent and reliable performance without requiring fine-tuning of the ensemble ratio.

F. Train and Test Split

The division of data into training and testing sets is a critical step in the development and evaluation of machine learning models [60]. The primary purpose of data splitting is to evaluate the performance and generalisation capacity of machine learning models. It involves segregating the data set into two distinct subsets: the training set and the testing set. The training set is utilised to train the model. It is usually a substantial portion of the data, allowing the model to learn patterns, relationships, and features from the input data [60]. The testing set serves as an independent dataset used to assess the model's performance. It evaluates how well the model can generalise its learned knowledge to unseen data [60].

The ratio at which data is split between the training and testing sets is a critical consideration. A common split ratio is 70/30 which has been adopted in this study, where 70% of the data is allocated to the training set, and the remaining 30% is assigned to the testing set.

V. Results

This section begins by analysing the results of the ablation study, leveraging its findings to conduct a more informed evaluation of class imbalance mitigation strategies. The study compares the performance of SMOTE, Cost-Sensitive Learning, and LEL using various base classifiers, including Random Forest (RF), Naive Bayes (NB), Decision Trees (DT), and XGBoost, across multiple datasets. Key performance metrics such as recall and precision are evaluated, with statistical significance assessed through paired t-tests and the Wilcoxon signed-rank test.

A. LELś Performance

Table IV summarizes the recall and precision values achieved by Random Forest (RF), Naive Bayes (NB), Decision Trees (DT), and XGBoost across different datasets. The variations tested include the base models, SMOTE, CSL and LEL. Among these techniques, LEL consistently outperformed the others, significantly enhancing predictive performance across all models.

For Random Forest, LEL achieved recall values as high as 0.96 and precision often exceeding 0.99. Statistical significance tests in Table V confirm these improvements as significant, with the paired t-test showing p-values of 0.0068 and 0.0256 compared to the base model and CSL, respectively. The Wilcoxon signed-rank test further supports these findings with p-values below 0.01.

In Naive Bayes models, LEL proved effective, especially for datasets like ”PIMA” and ”HABERMAN,” where recall values reached 0.90 and 0.81, and precision hit 0.92 for the ”P3” dataset. Statistical tests show p-values below 0.05 for most datasets, indicating significant improvements with LEL over SMOTE and CSL, which showed less impactful improvements.

Decision Trees also saw notable performance gains using LEL, with recall values up to 0.89 and precision up to 0.88. The statistical significance tests (Table V) confirm these results, with p-values below 0.05 in both paired t-tests and Wilcoxon signed-rank tests, validating LEL's superiority over the base model, SMOTE, and CSL.

Finally, XGBoost demonstrated substantial improvements with LEL, achieving recall values of 0.93 and precision consistently above 0.90. As shown in Table V, LEL produced statistically significant performance gains, with p-values below 0.05 in both paired t-tests and Wilcoxon signed-rank tests, highlighting its effectiveness over alternative techniques.

1) Shapely Analysis: The core idea of SHAP (Shapley Additive Explanations) values is to quantify the contribution of each feature by considering all possible combinations of feature subsets. By calculating the average marginal contribution of a feature across different subsets, SHAP provides a clear understanding of how features impact on model predictions [61]. The SHAP diagrams for each class imbalance mitigation strategy for the random forest experiment (Base, SMOTE, CSL, and LEL) against the P3 dataset are shown in Figure 6 to Figure 9.

The BASE model's SHAP plot (Figure 6) shows that features such as AVGHr, LASTSBP, and LASTO2SAT are the most influential in predicting sepsis. The dominance of heart rate and blood pressure highlights their relevance as early indicators of sepsis. This model leverages these cardiovascular metrics, which aligns with clinical findings that increased heart rate and blood pressure variations are early signs of infection [62]. However, the relatively narrow range of influential features used by the base model may limit its performance when predicting more complex cases of sepsis.

The SHAP values for the SMOTE model (Figure 7) reflect a shift towards respiratory features, particularly LASTO2SAT and MINAST. SMOTE oversampling emphasizes under-represented classes by generating synthetic samples; this influence increases the power of respiratory indicators, which improves predictions for patients with early respiratory failure, a common complication in sepsis [63][64]. This allows the model to balance its sensitivity towards minority class examples, improving recall for sepsis detection.

In the CSL model (Figure 8), features such as AGE, LASTO2SAT, and LASTSBP are prominent. The inclusion of age as a key feature suggests that the model incorporates both acute and chronic risk factors, accounting for the increased sepsis risk in elderly patients [62]. CSL adjusts for the cost of misclassification, making it more sensitive to high-risk cases of sepsis, which is crucial in improving precision and overall diagnostic accuracy. This cost-sensitive approach highlights features that are highly predictive in high-cost errors (e.g., missing a sepsis diagnosis).

The LEL modelâs SHAP plot (Figure 9) presents a more balanced feature importance distribution, with AVGTemp, LASTGlucose, and LASTPlatelets showing strong contributions. The use of an ensemble approach in LEL allows the model to draw from multiple classifiers, leading to a more robust feature set and improved predictive performance. This model accounts for a broader range of physiological markers, ensuring that key sepsis indicators are not overlooked [62]. For instance, the inclusion of temperature and glucose levels as critical features aligns with the understanding that metabolic and inflammatory responses play a significant role in sepsis pathophysiology.

A noteworthy observation across all SHAP diagrams is the consistent prominence of **temperature (AVGTemp)** as a key feature, regardless of the mitigation strategy employed. This highlights temperatureâs universal importance in predicting sepsis and underscores its role as a fundamental physiological indicator across different modelling approaches. The persistent significance of temperature suggests its critical role in capturing the inflammatory and metabolic responses associated with sepsis, making it indispensable for model interpretability and diagnostic accuracy.

One important feature to note is the generated attribute ”Sample Type” in the LEL SHAP analysis, which is the most discriminatory. This highlights its critical role in the model’s predictive capacity. As a generated attribute, it likely encapsulates information summarizing several underlying data points related to patient conditions and their clinical trajectories. Synthetic or derived features often encapsulate complex relationships between base features, providing better insights than individual raw features alone, which is evident in the ”Sample Type” feature [63][65][62].

The results generated in this study demonstrate that LEL outperforms the base classifier by leveraging a more comprehensive set of features and improving the handling of imbalanced data. By utilizing a robust combination of localized sampling strategies and enriched feature importance distributions, LEL achieves superior performance in sepsis prediction.

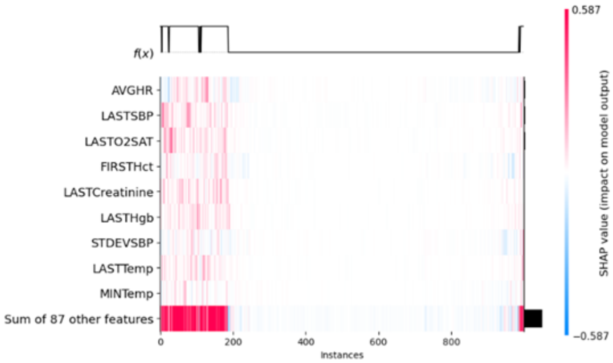

Fig. 6. SHAP plot for the Random Forest (RF) classifier on the P3 dataset without any imbalance mitigation strategy. The features AVGHr, LASTSBP, and LASTO2SAT are identified as the most influential in predicting sepsis. This reflects the importance of cardiovascular metrics such as heart rate and blood pressure in detecting early signs of sepsis. The narrow range of influential features in this model may limit its performance when handling more complex cases.

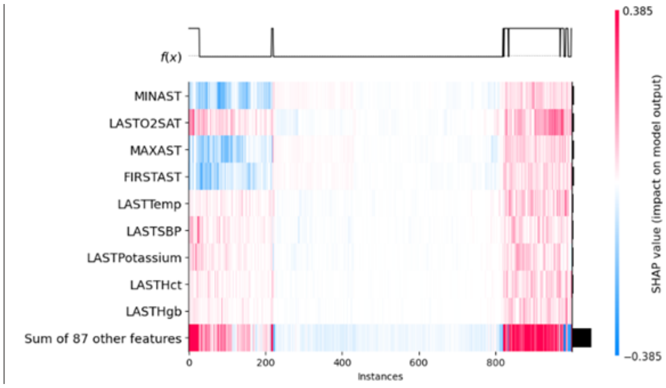

Fig. 7. SHAP plot for the Random Forest (RF) classifier with SMOTE applied on the P3 dataset. The feature importance shifts toward respiratory indicators, with LASTO2SAT and MINAST emerging as dominant predictors. This demonstrates SMOTE’s ability to enhance the model’s sensitivity to under-represented classes, improving recall for minority class examples such as patients with early respiratory complications in sepsis.

2) Precision-Recall Curve Analysis: The Precision-Recall (PR) curves, presented in Figures 10 to 13, illustrate the performance of three class imbalance mitigation strategies - SMOTE, Cost-Sensitive Learning (CSL), and Localized Ensemble Learning (LEL) â on the CONS dataset using Random Forest, Naive Bayes, Decision Trees, and XGBoost classifiers.

- **LEL Outperforms Alternative Methods:** Across all classifiers, the PR curve for LEL (green line) consistently dominates the curves for SMOTE and CSL. This demonstrates LEL’s ability to achieve a superior balance between precision and recall, which is critical for high-imbalance datasets like CONS.
- **Random Forest (Fig. 10):** LEL maintains high precision across nearly the entire recall range, significantly outper-

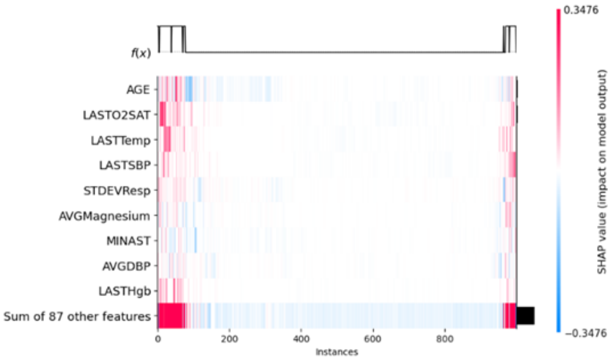

Fig. 8. SHAP plot for the Random Forest (RF) classifier with Cost-Sensitive Learning (CSL) applied on the P3 dataset. Features such as AGE, LASTO2SAT, and LASTSBP are prominent, indicating the model's consideration of both acute and chronic risk factors, including the increased sepsis risk in elderly patients. The CSL approach highlights features that minimize high-cost errors, such as missing sepsis diagnoses, thereby improving precision and diagnostic accuracy.

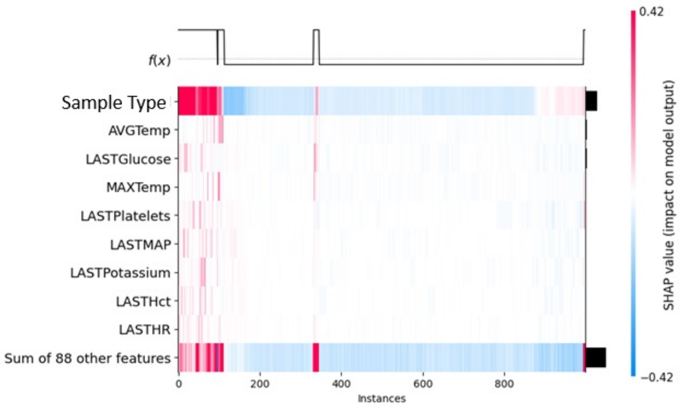

Fig. 9. SHAP plot for the Random Forest (RF) classifier with Localized Ensemble Learning (LEL) applied on the P3 dataset. The plot reveals a balanced feature importance distribution, with AVGTemp, LASTGlucose, and LASTPlatelets making significant contributions. This reflects LELâs ability to integrate a diverse range of physiological markers through localized sampling strategies and ensemble learning. The *Sample Type* feature, which encapsulates information about local neighborhood structures, emerges as the most discriminatory, underscoring its critical role in improving predictive performance.

forming CSL and SMOTE. Both SMOTE and CSL show a sharper decline in precision at higher recall values, indicating limitations in balancing false positives and false negatives effectively.

- **Naive Bayes (Fig. 11):** While LEL provides the best performance, the PR curve is less smooth compared to other classifiers, reflecting Naive Bayes' sensitivity to feature interactions and imbalanced datasets. SMOTE struggles significantly, with precision dropping steeply as recall increases, while CSL shows moderate performance but is still inferior to LEL.
- **Decision Trees (Fig. 12):** LEL demonstrates a near-optimal PR curve, consistently outperforming SMOTE and CSL across all recall values. CSL exhibits performance comparable to SMOTE at lower recall values but deteriorates significantly at higher recall levels.
- **XGBoost (Fig. 13):** Similar to Random Forest, XGBoost achieves excellent precision-recall tradeoffs when combined with LEL. Both SMOTE and CSL fail to maintain precision at higher recall values, underscoring their inability to handle complex imbalanced datasets as effectively as LEL.

VI. Discussion

In this section we discuss the results and highlight some interesting insights for both the ablation study and the actual experiements.

A. LEL

LEL's approach to addressing class imbalance at a granular level was pivotal to its success. By leveraging K-Nearest Neighbors (KNN) to classify samples into Safe, Borderline, and Rare categories, LEL applied targeted imbalance mitigation strategies tailored to each sample type. This localized treatment contrasts with traditional global methods, which uniformly apply a single strategy across all samples. By focusing on individual sample neighborhoods, LEL not only mitigated imbalance but also avoided pitfalls like overfitting or synthetic sample misplacement, common in global methods. This nuanced approach was particularly effective for datasets with complex minority class structures, such as PIMA and BUPA.

1) Dataset-Specific Performance: LELs versatility and adaptability were evident across datasets with varying levels of complexity and imbalance:

- **High-dimensional, complex datasets:** On PhysioNet-derived datasets (P3, P6, P12), LEL excelled in capturing temporal dependencies critical for sepsis prediction. For example, it achieved recall and precision of 0.96 and 0.99, respectively, on the P3 dataset using Random Forest, outperforming both SMOTE and CSL.
- **Moderately imbalanced datasets:** In simpler datasets like PIMA (IR = 1.86) and Haberman (IR = 2.7), where imbalance was less severe, LEL matched or outperformed traditional methods. For instance, LEL achieved recall of 0.70 and precision of 0.90 on the Haberman dataset for the Decision Tree classifier, significantly surpassing SMOTE and CSL.
- **Highly imbalanced datasets:** On extreme cases like CONS (IR = 20.56) and Poker (IR = 82), LELs granular treatment of minority instances enabled it to achieve perfect recall and precision in many cases, highlighting its robustness in handling severe imbalance.

These results underscore LELs ability to adapt to diverse data characteristics, whether in high-dimensional, temporally structured datasets or simpler, less imbalanced scenarios.

2) Classifier-Specific Analysis: LEL demonstrated its adaptability across classifiers:

- **Random Forest:** Achieved perfect recall and precision on CONS, highlighting its ability to generalize effectively.
- **Naive Bayes:** Showed unexpected improvements, such as a 56% increase in recall on BUPA, due to the inclusion of the ”Sample Type” feature.

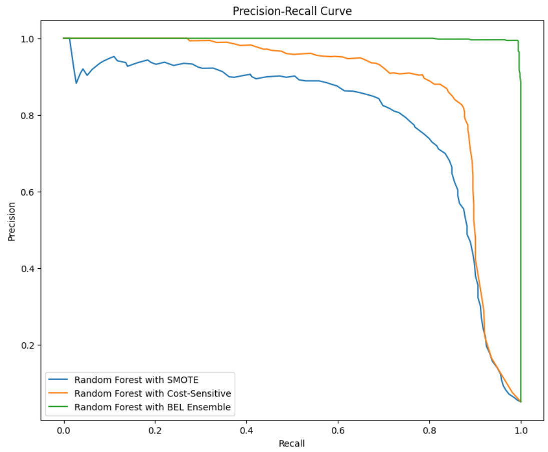

Fig. 10. Precision-Recall Curve for the CONS dataset using Random Forest. The curve compares three class imbalance mitigation strategies: SMOTE, Cost-Sensitive Learning (CSL), and Localized Ensemble Learning (LEL). LEL outperforms SMOTE and CSL by maintaining higher precision across nearly all recall values, demonstrating its ability to handle severe class imbalance effectively.

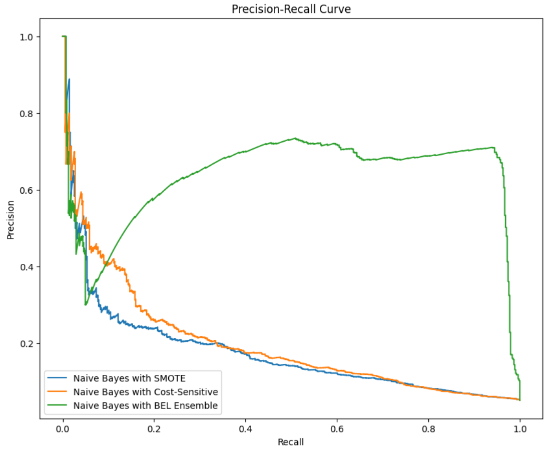

Fig. 11. Precision-Recall Curve for the CONS dataset using Naive Bayes. The PR curve for LEL shows significant improvements over SMOTE and CSL, although it exhibits more variability due to Naive Bayesâ sensitivity to noise and feature imbalances. SMOTE struggles with poor precision at higher recall values, while LEL demonstrates superior performance across the curve.

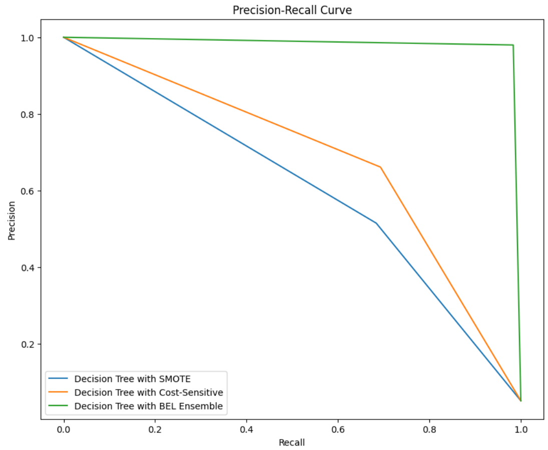

Fig. 12. Precision-Recall Curve for the CONS dataset using Decision Trees. LEL achieves a near-optimal PR curve, dominating SMOTE and CSL across all recall values. The results highlight LELs ability to improve decision boundary clarity and address class imbalance effectively, particularly in datasets with complex structures.

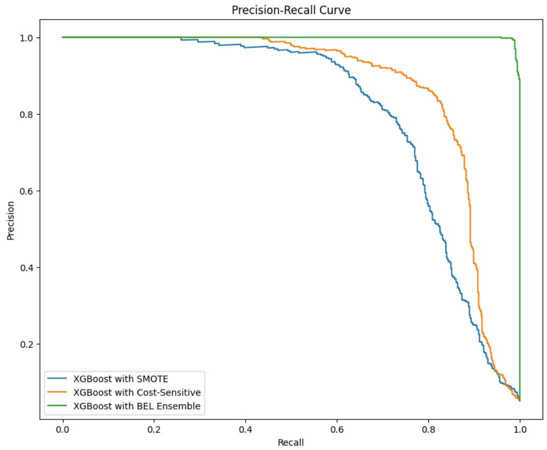

Fig. 13. Precision-Recall Curve for the CONS dataset using XGBoost. LEL provides excellent precision-recall tradeoffs, significantly outperforming SMOTE and CSL, particularly at higher recall values. The smoothness of the PR curve demonstrates XGBoosts ability to leverage LELs localized strategies effectively.

- **Decision Trees:** Benefited significantly from LELs localized approach, achieving consistent improvements across datasets.
- **XGBoost:** Enhanced recall and precision further demonstrating LELs robustness.

3) SHAP Analysis: The SHAP analysis revealed important insights into LELs feature utilization. Unlike traditional methods, LEL leveraged a broader range of features, with AVGTemp and LASTGlucose emerging as critical predictors for sepsis detection in PhysioNet datasets. Notably, the ”Sample Type” feature was the most discriminatory, encapsulating key information about minority and borderline instances. This highlights the value of synthetic or derived features in improving model performance.

4) Precision-Recall Curve Analysis: The analysis of PR curves highlights several key insights into the performance of the mitigation strategies:

- **LELs Superiority:** LEL outperforms SMOTE and CSL across all classifiers, demonstrating its ability to adaptively address class imbalance by leveraging localized mitigation strategies. LEL’s ability to maintain high precision across a wide range of recall values is critical for applications like healthcare, where both false positives and false negatives have severe consequences.
- **Classifier-Specific Trends:**
  - **Random Forest and XGBoost:** These classifiers exhibit smoother PR curves, indicating better utilization of LEL’s localized strategies. Their ensemble-based architectures amplify the benefits of LEL’s balanced feature importance distribution.
  - **Naive Bayes:** The variability in Naive Bayes’ PR curve reflects its sensitivity to noise and feature imbalances. While LEL improves its performance, the simpler probabilistic nature of Naive Bayes limits

its ability to fully leverage LEL's advantages.

  - **Decision Trees:** LEL significantly improves Decision Tree performance, achieving consistent gains in precision and recall. This highlights LEL's effectiveness in reducing overfitting and improving decision boundary clarity for simpler classifiers.

- **Limitations of SMOTE and CSL:** SMOTE and CSL exhibit suboptimal performance, particularly at higher recall values, as they fail to address nuances like neighborhood, distance, and quality imbalances within the dataset. Their global strategies are less effective compared to LELs localized approach.

B. Strengths and Contributions

LELs localized treatments significant advancements in addressing class imbalance. By dynamically adjusting treatments based on sample type, LEL achieved unparalleled performance across diverse datasets. Its adaptability to different classifiers and dataset characteristics demonstrates its potential for broad applicability in critical domains like healthcare and fraud detection.

C. Limitations and Future Directions

Despite its strengths, LEL has limitations:

- **Computational complexity:** Identifying sample types using KNN is resource-intensive, particularly for large datasets.
- **Threshold sensitivity:** Fixed thresholds for Safe, Rare, and Borderline classifications may limit performance and require further optimization.

Future work should explore adaptive thresholding and advanced neighborhood-based methods, such as density-based clustering, to enhance LELs scalability and performance. Integrating LEL with deep learning models could also address class imbalance in complex domains like image recognition and natural language processing.

VII. Conclusion

This research highlights the significant advantages of the Localised Ensemble Learning (LEL) approach in addressing class imbalance in machine learning. By adopting a localized and tailored strategy that focuses on the unique characteristics of each sample type Safe, Borderline, and Rare, LEL has demonstrated its effectiveness across multiple classifiers, including Random Forest, Naive Bayes, Decision Trees, and XGBoost. Experimental results consistently show that LEL outperforms traditional methods like SMOTE and Cost-Sensitive Learning (CSL) in recall and precision. These improvements were validated through statistical significance testing using paired t-tests and Wilcoxon signed-rank tests.

A key contribution of this work is the introduction of the ”Sample Type” feature, which played a pivotal role in enhancing model performance. As demonstrated by SHAP (SHapley Additive exPlanations) analysis, this feature enabled LEL to capture and leverage local data patterns, leading to more nuanced and accurate predictions, particularly for minority class instances.

The findings have both theoretical and practical implications. Theoretically, LEL advances the field by moving beyond global correction strategies that often fail to account for the complexities of localized data distributions. Practically, the approach has immediate relevance in high-stakes domains such as healthcare and fraud detection, where accurately identifying minority class instances can be life-saving or financially critical.

Despite its strengths, LEL is not without limitations. The methodâs reliance on neighbourhood-based sample classification introduces computational complexity, particularly for large-scale datasets. Further work is needed to optimize LEL for scalability. Additionally, while LEL demonstrated significant improvements over existing methods, its generalizability should be tested across a broader range of datasets and application domains.

Future research directions include refining LEL by exploring more sophisticated algorithms for determining sample types and extending its applicability to other complex learning problems. Investigating the integration of LEL with advanced techniques like deep learning could further enhance its utility in addressing class imbalance, particularly in domains requiring highly complex models.

References

[1] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[2] V. García, R. A. Mollineda, and J. S. Sánchez, “On the k-nn performance in a challenging scenario of imbalance and overlapping,” *Pattern Analysis and Applications*, vol. 11, no. 3-4, pp. 269–280, 2008.

[3] H. Ali, M. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq, “Imbalance class problems in data mining: A review,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 3, pp. 1560–1571, 2019.

[4] C. Arun and C. Lakshmi, “Diversity based multi-cluster over sampling approach to alleviate the class imbalance problem in software defect prediction,” *International Journal of System Assurance Engineering and Management*, pp. 1–13, 2023.

[5] J. M. N. Gøttcke and A. Zimek, “Handling class imbalance in k-nearest neighbor classification by balancing prior probabilities,” in *Similarity Search and Applications: 14th International Conference, SISAP 2021, Dortmund, Germany, September 29–October 1, 2021, Proceedings 14*. Springer, 2021, pp. 247–261.

[6] P. P. Wagle and M. Manoj Kumar, “A comprehensive review on the issue of class imbalance in predictive modelling,” *Emerging Research in Computing, Information, Communication and Applications: Proceedings of ERCICA 2022*, pp. 557–576, 2022.

[7] Y.-J. Park and K.-Y. Cheng, “A cluster impurity-based hybrid resampling for imbalanced classification problems,” *Applied Intelligence*, vol. 54, no. 20, pp. 9671–9684, 2024.

[8] M. Kubat, S. Matwin *et al.*, “Addressing the curse of imbalanced training sets: one-sided selection,” in *Icml*, vol. 97, no. 1. Citeseer, 1997, p. 179.

[9] S. Goswami and A. K. Singh, “A literature survey on various aspect of class imbalance problem in data mining,” *Multimedia Tools and Applications*, pp. 1–26, 2024.

[10] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. Chen, “A survey on imbalanced learning: latest research, applications and future directions,” *Artificial Intelligence Review*, vol. 57, no. 6, pp. 1–51, 2024.

[11] S. Rezvani and X. Wang, “A broad review on class imbalance learning techniques,” *Applied Soft Computing*, vol. 143, p. 110415, 2023.

[12] K. M. Hasib, M. S. Iqbal, F. M. Shah, J. A. Mahmud, M. H. Popel, M. I. H. Showrov, S. Ahmed, and O. Rahman, “A survey of methods for managing the classification and solution of data imbalance problem,” *arXiv preprint arXiv:2012.11870*, 2020.

[13] S. S. Rawat and A. K. Mishra, "Review of methods for handling class imbalance in classification problems," in *International Conference on Data, Engineering and Applications*. Springer, 2022, pp. 3–14.

[14] O. Wu, "Rethinking class imbalance in machine learning," *arXiv preprint arXiv:2305.03900*, 2023.

[15] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Machine Learning*, vol. 113, no. 7, pp. 4845–4901, 2024.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, pp. 321–357, 2002.

[17] Z. Liu, P. Wei, Z. Wei, B. Yu, J. Jiang, W. Cao, J. Bian, and Y. Chang, "Handling inter-class and intra-class imbalance in class-imbalanced learning," *arXiv preprint arXiv:2111.12791*, 2021.

[18] S. Das, S. S. Mullick, and I. Zelinka, "On supervised class-imbalanced learning: An updated perspective and some key challenges," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 6, pp. 973–993, 2022.

[19] S. Lusito, A. Pugnana, and R. Guidotti, "Solving imbalanced learning with outlier detection and features reduction," *Machine Learning*, vol. 113, no. 8, pp. 5273–5330, 2024.

[20] M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," *International Journal of Machine Learning and Computing*, vol. 3, p. 224, 04 2013.

[21] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing*. Springer, 2005, pp. 878–887.

[22] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, 2012.

[23] Y. Sun, H. Que, Q. Cai, J. Zhao, J. Li, Z. Kong, and S. Wang, "Borderline smote algorithm and feature selection-based network anomalies detection strategy," *Energies*, vol. 15, no. 13, p. 4751, 2022.

[24] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in *Advances in knowledge discovery and data mining: 13th Pacific-Asia conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 proceedings 13*. Springer, 2009, pp. 475–482.

[25] B. Liu, K. Blekas, and G. Tsoumakas, "Multi-label sampling based on local label imbalance," *Pattern Recognition*, vol. 122, p. 108294, 2022.

[26] Z. Teng, P. Cao, M. Huang, Z. Gao, and X. Wang, "Multi-label borderline oversampling technique," *Pattern Recognition*, vol. 145, p. 109953, 2024.

[27] K. Zhang, Z. Mao, P. Cao, W. Liang, J. Yang, W. Li, and O. R. Zaiane, "Label correlation guided borderline oversampling for imbalanced multi-label data learning," *Knowledge-Based Systems*, vol. 279, p. 110938, 2023.

[28] J. Liu, K. Huang, C. Chen, and J. Mao, "An oversampling algorithm of multi-label data based on cluster-specific samples and fuzzy rough set theory," *Complex & Intelligent Systems*, pp. 1–16, 2024.

[29] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (smote) for imbalanced learning," *Machine Learning*, vol. 113, no. 7, pp. 4903–4923, 2024.

[30] H. R. Sayegh, W. Dong, and A. M. Al-madani, "Enhanced intrusion detection with lstm-based model, feature selection, and smote for imbalanced data," *Applied Sciences*, vol. 14, no. 2, p. 479, 2024.

[31] U. Hasanah, A. M. Soleh, and K. Sadik, "Effect of random under sampling, oversampling, and smote on the performance of cardiovascular disease prediction models," *Jurnal Matematika, Statistika dan Komputasi*, vol. 21, no. 1, pp. 88–102, 2024.

[32] R. Blagus and L. Lusa, "Smote for high-dimensional class-imbalanced data," *BMC bioinformatics*, vol. 14, pp. 1–16, 2013.

[33] C. Yang, E. A. Fridgeirsson, J. A. Kors, J. M. Reps, and P. R. Rijnbeek, "Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data," *Journal of big data*, vol. 11, no. 1, p. 7, 2024.

[34] Q. Leng, J. Guo, J. Tao, X. Meng, and C. Wang, "Obmi: oversampling borderline minority instances by a two-stage tomek link-finding procedure for class imbalance problem," *Complex & Intelligent Systems*, pp. 1–18, 2024.

[35] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[36] F. Hutter, L. Kotthoff, and J. Vanschoren, *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.

[37] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of intelligent information systems*, vol. 17, pp. 107–145, 2001.

[38] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of data science*, vol. 2, pp. 165–193, 2015.

[39] V. C. Nitesh, "Smote: synthetic minority over-sampling technique," *J Artif Intell Res*, vol. 16, no. 1, p. 321, 2002.

[40] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. Ieee, 2008, pp. 1322–1328.

[41] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[42] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[43] M. A. Reyna, C. Josef, S. Seyedi, R. Jeter, S. P. Shashikumar, M. B. Westover, A. Sharma, S. Nemati, and G. D. Clifford, "Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019," in *2019 Computing in Cardiology (CinC)*. IEEE, 2019, pp. Page–1.

[44] J. Derrac, S. Garcia, L. Sanchez, and F. Herrera, "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *J. Mult. Valued Logic Soft Comput*, vol. 17, pp. 255–287, 2015.

[45] F. Kamran, D. Tjandra, A. Heiler, J. Virzi, K. Singh, J. E. King, T. S. Valley, and J. Wiens, "Evaluation of sepsis prediction models before onset of treatment," *NEJM AI*, vol. 1, no. 3, p. AIoa2300032, 2024.

[46] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 801–810, 2016.

[47] V. Braverman, R. Ostrovsky, and C. Zaniolo, "Optimal sampling from sliding windows," in *Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2009, pp. 147–156.

[48] B. Mwangi, T. S. Tian, and J. C. Soares, "A review of feature reduction techniques in neuroimaging," *Neuroinformatics*, vol. 12, pp. 229–244, 2014.

[49] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *2014 science and information conference*. IEEE, 2014, pp. 372–378.

[50] J. Peng, J. Hahn, and K.-W. Huang, "Handling missing values in information systems research: A review of methods and assumptions," *Information Systems Research*, vol. 34, no. 1, pp. 5–26, 2023.

[51] J.-M. Jo, "Effectiveness of normalization pre-processing of big data to the machine learning performance," *The Journal of the Korea institute of electronic communication sciences*, vol. 14, no. 3, pp. 547–552, 2019.

[52] V. Priyalakshmi and R. Devi, "Analysis and implementation of normalisation techniques on kddâ99 data set for ids and ips," in *Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 2*. Springer, 2023, pp. 51–70.

[53] A. S. Saif, A. G. Garba, J. Awwalu, H. Arshad, and L. Q. Zakaria, "Performance comparison of min-max normalisation on frontal face detection using haar classifiers," *Pertanika J. Sci. Technol*, vol. 25, pp. 163–171, 2017.

[54] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[55] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[56] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, pp. 81–106, 1986.

[57] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[58] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[59] N. Japkowicz, "Assessment metrics for imbalanced learning," *Imbalanced learning: Foundations, algorithms, and applications*, pp. 187–206, 2013.

[60] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, no. 1, pp. 381–386, 2020.

[61] J. Hirsch, R. L. DeLaPaz, N. R. Relkin, J. Victor, K. Kim, T. Li, P. Borden, N. Rubin, and R. Shapley, "Illusory contours activate specific regions in human visual cortex: evidence from functional magnetic resonance imaging." *Proceedings of the National Academy of Sciences*, vol. 92, no. 14, pp. 6469–6473, 1995.

[62] U. Aygun, F. H. Yagin, B. Yagin, S. Yasar, C. Colak, A. S. Ozkan, and L. P. Ardigò, "Assessment of sepsis risk at admission to the emergency department: Clinical interpretable prediction model," *Diagnostics*, vol. 14, no. 5, p. 457, 2024.

[63] S. Bomrah, M. Uddin, U. Upadhyay, M. Komorowski, J. Priya, E. Dhar, S.-C. Hsu, and S. Syed-Abdul, "A scoping review of machine learning for sepsis prediction-feature engineering strategies and model performance: a step towards explainability," *Critical Care*, vol. 28, no. 1, p. 180, 2024.

[64] M. S. Rahman, K. R. Islam, J. Prithula, J. Kumar, M. Mahmud, M. F. Alam, M. B. I. Reaz, A. Alqahtani, and M. E. Chowdhury, "Machine learning-based prognostic model for 30-day mortality prediction in sepsis-3," *BMC medical informatics and decision making*, vol. 24, no. 1, p. 249, 2024.

[65] E. S. Rangan, R. K. Pathinarupothi, K. J. Anand, and M. P. Snyder, "Performance effectiveness of vital parameter combinations for early warning of sepsisâan exhaustive study using machine learning," *JAMIA open*, vol. 5, no. 4, p. ooac080, 2022.

TABLE III

This table presents the performance metrics (Recall and Precision) for various experiments conducted in the ablation study, evaluating different class mitigation strategy configurations across multiple k-values. Each row corresponds to a specific experimental setup, detailing the applied strategies for the Safe, Borderline, and Rare sample types. The columns include: Experiment, which identifies the setup; K Value, indicating the number of nearest neighbors used; Safe, Borderline, Rare, which represent ratio for each Classifer in the Ensemble; Safe Strategy, Borderline Strategy, and Rare Strategy, specifying the class imbalance mitigation methods applied to each sample type (e.g., NearMiss, SMOTE, etc.); and Recall and Precision, the performance metrics reflecting the configuration's ability to correctly identify positive samples and the accuracy of positive predictions, respectively. The table allows a detailed comparison of the strategiesÂ effectiveness in mitigating class imbalance under varying parameter settings.

Experiment	K Value	Safe	Borderline	Rare	Safe Strategy	Borderline Strategy	Rare Strategy	Recall	Precision
1	5	6	1	1	NearMiss	Random Oversample	NearMiss	**0.4657**	**0.1306**
		1	6	1	NearMiss	Random Oversample	NearMiss	**0.4657**	**0.1306**
		1	1	6	NearMiss	Random Oversample	NearMiss	**0.4657**	**0.1306**
		1	1	1	NearMiss	Random Oversample	NearMiss	**0.4657**	**0.1306**
	10	6	1	1	NearMiss	Random Oversample	NearMiss	0.4584	0.1284
		1	6	1	NearMiss	Random Oversample	NearMiss	0.4584	0.1284
		1	1	6	NearMiss	Random Oversample	NearMiss	0.4584	0.1284
		1	1	1	NearMiss	Random Oversample	NearMiss	0.4584	0.1284
	15	6	1	1	NearMiss	Random Oversample	NearMiss	0.4584	0.1283
		1	6	1	NearMiss	Random Oversample	NearMiss	0.4584	0.1283
		1	1	6	NearMiss	Random Oversample	NearMiss	0.4584	0.1283
		1	1	1	NearMiss	Random Oversample	NearMiss	0.4584	0.1283
	20	6	1	1	NearMiss	Random Oversample	NearMiss	0.4561	0.1276
		1	6	1	NearMiss	Random Oversample	NearMiss	0.4561	0.1276
		1	1	6	NearMiss	Random Oversample	NearMiss	0.4561	0.1276
		1	1	1	NearMiss	Random Oversample	NearMiss	0.4561	0.1276
2	5	6	1	1	SMOTE	Random Oversample	TomekLinks	0.8736	0.8661
		1	6	1	SMOTE	Random Oversample	TomekLinks	0.8736	0.8661
		1	1	6	SMOTE	Random Oversample	TomekLinks	0.8736	0.8661
		1	1	1	SMOTE	Random Oversample	TomekLinks	0.8736	0.8661
	10	6	1	1	SMOTE	Random Oversample	TomekLinks	**0.8739**	**0.8670**
		1	6	1	SMOTE	Random Oversample	TomekLinks	**0.8739**	**0.8670**
		1	1	6	SMOTE	Random Oversample	TomekLinks	**0.8739**	**0.8670**
		1	1	1	SMOTE	Random Oversample	TomekLinks	**0.8739**	**0.8670**
	15	6	1	1	SMOTE	Random Oversample	TomekLinks	0.8739	0.8669
		1	6	1	SMOTE	Random Oversample	TomekLinks	0.8739	0.8669
		1	1	6	SMOTE	Random Oversample	TomekLinks	0.8739	0.8669
		1	1	1	SMOTE	Random Oversample	TomekLinks	0.8739	0.8669
	20	6	1	1	SMOTE	Random Oversample	TomekLinks	0.8739	0.8669
		1	6	1	SMOTE	Random Oversample	TomekLinks	0.8739	0.8669
		1	1	6	SMOTE	Random Oversample	TomekLinks	0.8739	0.8669
		1	1	1	SMOTE	Random Oversample	TomekLinks	0.8739	0.8669
3	5	6	1	1	SMOTEENN	CondensedNearestNeighbour	Random Oversample	**0.8719**	0.8549
		1	6	1	SMOTEENN	CondensedNearestNeighbour	Random Oversample	**0.8719**	0.8549
		1	1	6	SMOTEENN	CondensedNearestNeighbour	Random Oversample	**0.8719**	0.8549
		1	1	1	SMOTEENN	CondensedNearestNeighbour	Random Oversample	**0.8719**	0.8549
	10	6	1	1	SMOTEENN	CondensedNearestNeighbour	Random Oversample	0.8671	**0.8157**
		1	6	1	SMOTEENN	CondensedNearestNeighbour	Random Oversample	0.8671	**0.8157**
		1	1	6	SMOTEENN	CondensedNearestNeighbour	Random Oversample	0.8671	**0.8157**
		1	1	1	SMOTEENN	CondensedNearestNeighbour	Random Oversample	0.8671	**0.8157**
	15	6	1	1	SMOTEENN	CondensedNearestNeighbour	Random Oversample	0.8648	0.7874
		1	6	1	SMOTEENN	CondensedNearestNeighbour	Random Oversample	0.8648	0.7874
		1	1	6	SMOTEENN	CondensedNearestNeighbour	Random Oversample	0.8648	0.7874
		1	1	1	SMOTEENN	CondensedNearestNeighbour	Random Oversample	0.8648	0.7874
	20	6	1	1	SMOTEENN	CondensedNearestNeighbour	Random Oversample	0.8582	0.6860
		1	6	1	SMOTEENN	CondensedNearestNeighbour	Random Oversample	0.8582	0.6860
		1	1	6	SMOTEENN	CondensedNearestNeighbour	Random Oversample	0.8582	0.6860
		1	1	1	SMOTEENN	CondensedNearestNeighbour	Random Oversample	0.8582	0.6860
4	5	6	1	1	CondensedNearestNeighbour	SMOTETomek	ADASYN	**0.8720**	**0.8603**
		1	6	1	CondensedNearestNeighbour	SMOTETomek	ADASYN	**0.8720**	**0.8603**
		1	1	6	CondensedNearestNeighbour	SMOTETomek	ADASYN	**0.8720**	**0.8603**
		1	1	1	CondensedNearestNeighbour	SMOTETomek	ADASYN	**0.8720**	**0.8603**
	10	6	1	1	CondensedNearestNeighbour	SMOTETomek	ADASYN	0.8597	0.7923
		1	6	1	CondensedNearestNeighbour	SMOTETomek	ADASYN	0.8597	0.7923
		1	1	6	CondensedNearestNeighbour	SMOTETomek	ADASYN	0.8597	0.7923
		1	1	1	CondensedNearestNeighbour	SMOTETomek	ADASYN	0.8597	0.7923
	15	6	1	1	CondensedNearestNeighbour	SMOTETomek	ADASYN	0.8522	0.6957
		1	6	1	CondensedNearestNeighbour	SMOTETomek	ADASYN	0.8522	0.6957
		1	1	6	CondensedNearestNeighbour	SMOTETomek	ADASYN	0.8522	0.6957
		1	1	1	CondensedNearestNeighbour	SMOTETomek	ADASYN	0.8522	0.6957
	20	6	1	1	CondensedNearestNeighbour	SMOTETomek	ADASYN	0.8520	0.6750
		1	6	1	CondensedNearestNeighbour	SMOTETomek	ADASYN	0.8520	0.6750
		1	1	6	CondensedNearestNeighbour	SMOTETomek	ADASYN	0.8520	0.6750
		1	1	1	CondensedNearestNeighbour	SMOTETomek	ADASYN	0.8520	0.6750

TABLE IV
RECALL AND PRECISION RESULTS ACROSS DIFFERENT MODELS AND DATASETS

Model	**Dataset**	**RECALL**				**PRECISION**			
		Base	**Smote**	**CSL**	**LEL**	**Base**	**Smote**	**CSL**	**LEL**
Random Forest	P3	0.85	0.91	0.82	**0.96****	0.98	0.95	0.98	**0.99****
	P6	0.75	0.82	0.82	**0.86****	0.75	0.84	0.84	**0.97****
	P12	0.85	**0.97****	0.96	0.95	0.85	0.96	**0.97****	**0.97****
	HABERMAN	0.65	0.60	0.11	**0.70****	0.72	0.78	0.33	**0.90****
	BUPA	0.65	0.33	0.72	**1.00****	0.72	0.20	0.74	**1.00****
	POKER	0.54	0.33	0.00	**1.00****	0.00	0.20	0.74	**1.00****
	PIMA	0.67	0.33	0.59	**0.80****	0.72	0.20	0.70	**0.87****
	CONS	0.63	0.66	0.59	**1.00****	0.96	0.85	0.96	**1.00****
Naive Bayes	P3	0.85	0.85	0.88	**0.90****	0.87	0.87	0.77	**0.92****
	P6	0.70	0.58	0.86	**0.89****	0.80	0.80	0.73	**0.85****
	P12	0.69	0.66	**0.82****	**0.82****	0.75	0.76	0.67	**0.77****
	HABERMAN	0.74	0.71	0.71	**0.81****	0.74	0.68	0.50	**0.88****
	BUPA	0.62	0.57	0.43	**0.89****	0.70	0.67	0.61	**0.85****
	POKER	0.54	0.73	0.75	**0.99****	0.23	0.23	0.29	**0.89****
	PIMA	0.73	0.72	0.73	**0.78****	0.24	0.32	0.29	**0.79****
	CONS	0.74	0.75	0.74	**0.89****	0.10	0.10	0.10	**0.70****
Decision Trees	P3	**0.96**	0.94	0.88	**0.96****	**0.96**	0.94	0.77	**1.00****
	P6	**0.95**	0.88	0.85	**1.00****	**0.95**	0.89	0.72	**1.00****
	P12	**0.96**	0.89	0.82	**1.00****	**0.96**	0.89	0.67	**1.00****
	HABERMAN	0.68	0.68	**0.71****	0.68	0.66	0.62	0.50	**0.70****
	BUPA	0.67	0.64	0.67	**0.81****	0.69	0.68	0.69	**0.74****
	POKER	**0.98**	0.96	**1.00****	**1.00****	0.99	**1.00****	**1.00****	**1.00****
	PIMA	0.75	0.77	0.62	**0.82****	0.77	0.77	0.40	**0.88****
	CONS	0.97	0.97	0.95	**1.00****	0.97	0.97	0.96	**1.00****
XGBoost	P3	**0.94**	0.93	0.88	**0.96****	**0.94**	0.92	0.77	**0.98****
	P6	**0.91**	0.88	0.85	**0.93****	**0.91**	0.87	0.72	**0.92****
	P12	**0.90**	0.88	0.81	**0.95****	**0.88**	0.87	0.65	**0.93****
	HABERMAN	0.73	0.63	0.71	**0.80****	0.70	0.61	0.50	**0.75****
	BUPA	**0.77**	0.72	0.67	**0.87****	**0.78**	0.76	0.73	**0.85****
	POKER	**1.00****	0.98	**1.00****	**1.00****	**1.00****	1.00*	**1.00****	**1.00****
	PIMA	**0.76**	0.75	0.63	**0.85****	**0.76**	0.77	0.40	**0.80****
	CONS	**0.98**	0.98	0.95	**1.00****	0.95	**0.98**	0.90	**1.00****

TABLE V
STATISTICAL TEST SIGNIFICANCE FOR DIFFERENT MODELS' PRECISION AND ACCURACY COMPARED TO LEL

		RECALL			**PRECISION**		
Model	**Statistic**	**Base/LEL**	**Smote/LEL**	**CSL/LEL**	**Base/LEL**	**Smote/LEL**	**CSL/LEL**
Random Forest	Pair T-Test	**YES**	NO	**YES**	NO	**YES**	NO
	Wilcoxon	**YES**	NO	**YES**	**YES**	**YES**	**YES**
Naive Bayes	Pair T-Test	**YES**	**YES**	NO	**YES**	**YES**	**YES**
	Wilcoxon	**YES**	**YES**	NO	**YES**	**YES**	**YES**
Decision Trees	Pair T-Test	NO	**YES**	**YES**	**YES**	**YES**	**YES**
	Wilcoxon	NO	**YES**	**YES**	**YES**	**YES**	**YES**
XGBoost	Pair T-Test	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**
	Wilcoxon	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**

Note: ***YES*** indicates that LEL improved prediction more than the method in comparison with confidence level $\alpha >= 0.95$.