ISSN: 2473-4901
v12 n6841

Improving Clinical Decision Support for Sepsis Mortality Prediction Using Data Balancing Techniques

Dara Tourt
dara.tourt@my.metrostate.edu
Metro State University
Saint Paul, MN 55106, USA

Queen E. Booker
queen.booker@metrostate.edu
Metro State University
Saint Paul, MN 55106, USA

Carl M. Rebman, Jr.
carlr@sandiego.edu
University of San Diego
San Diego, CA 92110, USA

Abstract

Healthcare organizations increasingly rely on predictive analytics to support clinical decision making. However, severe class imbalance often limits the ability of predictive models to identify patients at greatest risk of mortality. This study examines the impact of commonly used data balancing techniques on the performance of an interpretable Logistic Regression model using the Sepsis Survival Minimal Clinical Records dataset containing over 110,000 ICU admissions. Results demonstrate that data balancing substantially improves mortality detection compared with an unbalanced baseline, although improvements in recall are accompanied by reductions in precision. The findings illustrate the trade-offs healthcare organizations must consider when implementing predictive analytics within clinical decision support systems and provide practical recommendations for balancing sensitivity, alert burden, and model transparency.

Keywords: Clinical Decision Support Systems, Sepsis Mortality Prediction, Predictive Analytics, Class Imbalance, Logistic Regression, Healthcare Information Systems

ISSN: 2473-4901
v12 n6841

Improving Clinical Decision Support for Sepsis Mortality Prediction Using Data Balancing Techniques

Dara Tourt, Queen E. Booker and Carl M Redman, Jr.

I. Introduction

Healthcare organizations are becoming increasingly reliant on predictive analytics and clinical decision support systems (CDSS) to assist medical professionals in identifying patients at elevated risk for adverse outcomes. Sepsis is among the most significant causes of hospital mortality worldwide. The Global Burden of Disease Study estimated approximately 48.9 million cases of sepsis and 11 million sepsis-related deaths annually, accounting for nearly one in five deaths worldwide (Rudd et al., 2020).

Early identification of patients at risk is therefore an important objective of modern clinical decision support, because timely and clinically relevant information can help providers prioritize interventions and allocate scarce resources. However, the value of a CDSS depends not only on predictive discrimination but also on whether its output is sufficiently reliable, interpretable, and appropriately integrated into clinical work (Sutton et al., 2020; Bayor et al., 2025). Excessive or low-value alerts can increase workload and contribute to alert fatigue, reducing the practical usefulness of decision support (Ancker et al., 2017).

Although advances in machine learning have improved predictive capabilities in many clinical applications, implementing these models within operational healthcare environments remains challenging. Clinical decision support must provide predictions that are not only statistically accurate but also understandable and usable within existing workflows. Research on decision support acceptance identifies usefulness, workflow and integration, ease of use, and trust as important adoption considerations (Shibl et al., 2013). In this context, simpler and more interpretable models such as Logistic Regression remain attractive because their behavior can be examined more readily than that of complex black-box algorithms.

One of the greatest obstacles to developing effective predictive models for sepsis mortality is the highly imbalanced nature of clinical data. Most patients survive, while relatively few experience mortality, resulting in datasets in which the minority class may represent less than ten percent of all observations. Standard classification algorithms trained on these data often achieve high overall accuracy while failing to identify many of the patients most in need of clinical intervention. From a clinical perspective, these false negatives are particularly problematic because missed high-risk patients represent lost opportunities for timely treatment.

A variety of data balancing techniques have been proposed to address this challenge by improving the representation of minority cases during model development. While previous studies have extensively compared balancing techniques across multiple datasets and machine learning algorithms, relatively little attention has been given to how these approaches influence the practical usefulness of predictive models within clinical decision support systems. Healthcare organizations must balance competing objectives: maximizing the identification of patients at risk while minimizing unnecessary clinical alerts that contribute to alarm fatigue and inefficient resource utilization.

This study presents an applied evaluation of data balancing techniques using the Sepsis Survival Minimal Clinical Records dataset, a large real-world dataset containing more than 110,000 intensive care unit admissions with a severe class imbalance of approximately 93% survival and 7% mortality. Rather than claiming algorithmic novelty or evaluating an operational CDSS, this research examines a design-stage analytics question: how established resampling choices change the performance profile of an interpretable Logistic Regression model that could supply risk information to clinical decision support. Model performance is evaluated using Recall, Precision, F1-score, ROC-AUC, and Precision-Recall AUC, with particular attention to how the choice of metric changes the apparent usefulness of the same predictive model.

This study makes two primary contributions. First, it demonstrates that the choice of both data-balancing strategy and evaluation metric can lead to substantially different assessments of the same predictive

model: methods that appear nearly unchanged under ROC-AUC can differ dramatically in their ability to identify mortality cases. Second, it frames those differences as a clinical decision-support design issue. Preprocessing choices affect the information a predictive component supplies to clinicians, while evaluation choices affect which performance trade-offs system designers see and prioritize. The contribution is therefore not a new resampling algorithm, but evidence about how established analytics choices shape the design and evaluation of interpretable predictive decision support.

Accordingly, this study addresses two research questions:
RQ1: How does data balancing affect the performance of Logistic Regression for mortality classification in a severely imbalanced sepsis dataset?
RQ2: How do alternative evaluation metrics characterize the performance trade-offs produced by different data balancing techniques?

The remainder of this paper is organized as follows. Section 2 reviews prior research on clinical decision support systems, predictive modeling for sepsis mortality, class imbalance, and the information-systems implications of predictive model design. Section 3 describes the methodology, dataset, preprocessing procedures, and evaluation framework. Section 4 presents the empirical findings and discusses their implications for clinical decision support. Section 5 identifies limitations and directions for future research.

2. Literature Review

2.1 Clinical Decision Support Systems for Sepsis

Clinical Decision Support Systems (CDSS) have become an integral component of modern healthcare information systems, assisting clinicians in making timely, evidence-based decisions by combining patient data with predictive analytics and clinical knowledge (Berner, 2007). Advances in electronic health records (EHRs), data integration, and predictive modeling have enabled healthcare organizations to incorporate decision support tools directly into clinical workflows, improving diagnostic accuracy, treatment decisions, and patient outcomes (Sutton et al., 2020).

One of the most promising applications of clinical decision support is the early identification of patients with sepsis. Sepsis is a life-threatening condition resulting from a dysregulated response to infection and remains one of the leading causes of mortality worldwide. The Global Burden of Disease Study estimated approximately 48.9 million sepsis cases and 11 million sepsis-related deaths annually, accounting for nearly one in five deaths globally (Rudd et al., 2020). Because early intervention substantially improves survival, healthcare organizations increasingly rely on predictive analytics to identify patients at elevated risk before clinical deterioration occurs.

Despite advances in predictive modeling, successful implementation of clinical decision support extends beyond statistical performance. Predictive systems must produce information that clinicians can trust and incorporate into existing workflows. Decision-support acceptance research identifies usefulness, facilitating conditions such as workflow and integration, ease of use, and trust as important influences on adoption (Shibl et al., 2013). Similarly, recent CDSS design research identifies usability, workflow integration, data quality, validity, reliability, and explainability as recurring design challenges (Bayor et al., 2025). Excessive false alarms can contribute to alert fatigue and lower alert acceptance, reducing the effectiveness of decision support (Ancker et al., 2017). Consequently, predictive performance measures have an information-systems implication: the balance between missed high-risk cases and false-positive alerts affects the information burden presented to clinical users.

2.2 Predictive Analytics for Sepsis Mortality

Machine learning has become an increasingly important tool for predicting patient outcomes, including mortality among patients diagnosed with sepsis. Numerous studies have demonstrated that predictive models can identify high-risk patients earlier than traditional rule-based clinical scoring systems, enabling more timely intervention and improved allocation of healthcare resources (Shashikumar et al., 2017; Komorowski et al., 2018).

While recent research has explored increasingly sophisticated algorithms, including random forests, gradient boosting, deep neural networks, and ensemble learning, these approaches often sacrifice

interpretability for incremental improvements in predictive accuracy. In healthcare environments, explainability remains a critical requirement because clinicians are more likely to trust and adopt prediction models whose behavior can be understood and justified.

Interestingly, systematic reviews have questioned whether increasingly complex machine learning models consistently outperform traditional statistical approaches. Christodoulou et al. (2019) found little evidence that machine learning algorithms consistently provide superior predictive performance compared with Logistic Regression across clinical prediction tasks. These findings suggest that interpretable models remain viable alternatives, particularly when transparency, reproducibility, and clinician trust are organizational priorities.

Logistic Regression continues to be widely used in healthcare because it produces probabilistic predictions that clinicians can readily interpret while maintaining relatively strong predictive performance. Consequently, Logistic Regression remains an appropriate foundation for clinical decision support systems intended for operational deployment.

2.3 Class Imbalance in Clinical Prediction

Although predictive modeling has demonstrated considerable promise in healthcare, many clinical datasets exhibit substantial class imbalance. Mortality, disease recurrence, adverse drug reactions, and rare disease diagnoses typically occur far less frequently than non-events, creating datasets in which the minority class may comprise only a small percentage of observations.

Class imbalance presents significant challenges for conventional classification algorithms because models naturally optimize overall accuracy by favoring the majority class (He & Garcia, 2009). As a result, models may achieve impressive accuracy while failing to identify the relatively few patients experiencing the clinically significant outcome. In sepsis mortality prediction, this translates into false negatives—patients who ultimately die but are classified as low risk. Such errors may prevent clinicians from initiating timely interventions, thereby reducing the practical value of predictive analytics.

Traditional performance measures can further obscure these deficiencies. Receiver Operating Characteristic (ROC) curves frequently remain favorable even when minority class detection is poor because false positive rates change relatively little in highly imbalanced datasets. In contrast, Precision-Recall (PR) curves provide a more informative assessment of classifier performance under severe imbalance by emphasizing the model's ability to correctly identify minority class observations (Saito & Rehmsmeier, 2015). Consequently, healthcare researchers increasingly recommend evaluating predictive models using Recall, Precision, F1-score, and PR-AUC in addition to overall accuracy and ROC-AUC.

2.4 Data Balancing Techniques for Healthcare Decision Support

To address class imbalance, researchers have developed numerous preprocessing techniques that modify the distribution of training data before model development. These approaches seek to improve the classifier's ability to identify minority class observations without fundamentally altering the underlying prediction algorithm.

Oversampling techniques increase the representation of minority class observations by duplicating existing cases or generating synthetic examples. Among these methods, the Synthetic Minority Oversampling Technique (SMOTE) remains one of the most widely adopted approaches because it creates synthetic observations through interpolation between neighboring minority cases rather than simple duplication (Chawla et al., 2002). Subsequent variations, including Borderline-SMOTE and ADASYN, seek to improve minority representation by focusing synthetic generation on difficult or poorly represented observations (Han et al., 2005; He et al., 2008).

Data cleaning techniques attempt to improve class separation by removing noisy or ambiguous observations near the decision boundary. Tomek Links eliminate overlapping majority observations, while Edited Nearest Neighbors (ENN) remove samples that are inconsistent with their local neighborhood (Tomek, 1976; Wilson, 1972). Hybrid approaches, including SMOTE combined with ENN or Tomek Links, seek to capitalize on the strengths of both oversampling and data cleaning by simultaneously improving minority representation while reducing class overlap (Batista et al., 2004).

Although these techniques have demonstrated improvements in recall across numerous healthcare applications, no single balancing strategy consistently outperforms others across all clinical datasets (Fernandez et al., 2018; Salmi et al., 2024). Moreover, improving minority detection frequently increases false positive rates, requiring healthcare organizations to carefully consider the operational consequences of increased clinical alerts.

2.5 Research Gap

Existing research has extensively evaluated machine learning algorithms and data balancing techniques for imbalanced healthcare datasets. However, much of this work emphasizes algorithmic performance, while information-systems research on decision support emphasizes a broader set of design concerns, including information quality, interpretability, workflow integration, trust, and user burden (Shibl et al., 2013; Miah et al., 2020; Bayor et al., 2025). The present study does not evaluate clinician acceptance or an operational CDSS. Instead, it addresses an earlier design-stage problem: whether preprocessing and evaluation choices materially change the information characteristics of a predictive component that may later be embedded in decision support.

Furthermore, many comparative studies evaluate multiple machine learning algorithms simultaneously, making it difficult to isolate the effect of data preparation on predictive performance. By holding the classifier constant and varying the balancing strategy, this study isolates how resampling affects the performance profile of an interpretable Logistic Regression model. This narrower design provides a controlled comparison of sensitivity, precision, and related metrics rather than a search for the most accurate algorithm. From a decision-support perspective, that comparison matters because false negatives and false positives translate into very different information and alert burdens for a hospital environment.

Building on the preceding literature, this study conceptualizes sepsis mortality prediction as a predictive component within a broader information-system design process. Clinical data are transformed through preprocessing and balancing choices; those choices shape model outputs; and model outputs, if incorporated into a CDSS, become information presented to clinicians. This framing is consistent with decision-support design research that treats the design of information artifacts and reusable design knowledge as central IS concerns (Miah et al., 2020). Figure 1 presents the conceptual framework guiding this study. The framework does not imply that clinician behavior or implementation outcomes are tested here; rather, it identifies the point at which analytics design choices can affect downstream decision-support information.

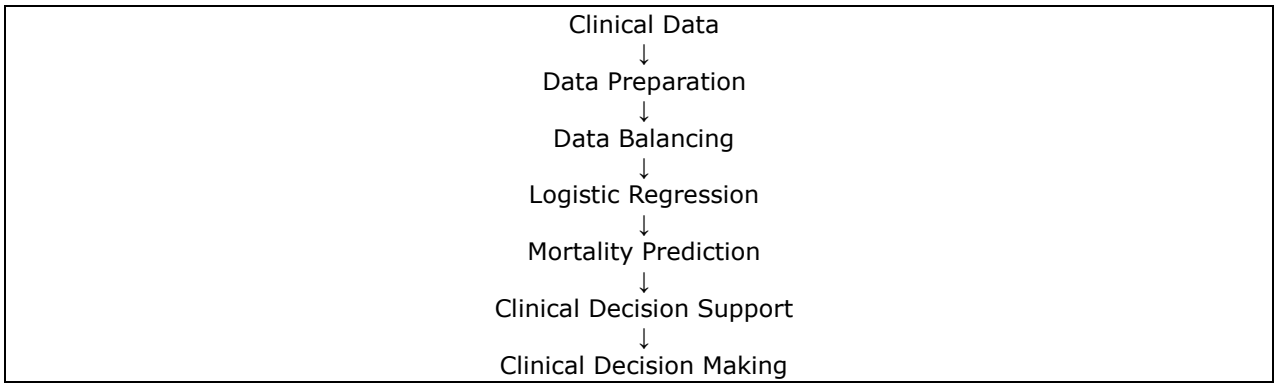

Figure 1. Conceptual Framework

3. Research Methodology

3.1 Dataset

This study utilizes the Sepsis Survival Minimal Clinical Records dataset obtained from the UCI Machine Learning Repository. The dataset contains 110,204 intensive care unit (ICU) admissions and includes

five routinely collected clinical variables used to predict patient survival following sepsis diagnosis. Approximately seven percent of patients in the dataset experienced mortality, producing a class distribution of approximately 93% survival and 7% mortality. This level of imbalance presents a significant challenge for predictive modeling because conventional classification algorithms tend to favor the majority class, often failing to identify patients at greatest clinical risk.

One of the strengths of this dataset is its reliance on a minimal number of routinely available clinical variables, making it representative of the types of information commonly available during early clinical assessments. Consequently, it provides an appropriate testbed for evaluating predictive analytics intended for operational clinical decision support.

3.2 Analysis Environment

All analyses were conducted in Python using the Google Colab cloud-based environment and scikit-learn/imbalanced-learn compatible pipelines. A fixed random state was used where applicable to support reproducibility. The analytical sequence was defined before resampling comparisons: stratified train/test partitioning, training-fold preprocessing and resampling, Logistic Regression estimation, cross-validation, and final evaluation on the held-out test set.

3.3 Data Preparation

To ensure consistency across datasets and prepare them for fair comparative analysis, a standardized data preprocessing pipeline was applied to the healthcare datasets. First, the dataset was examined for missing values. For variables with limited missingness, mean imputation was applied to numerical features and mode imputation to categorical variables.

3.4 Model Development

Following data preparation, the next phase involved model development and performance evaluation using a consistent experimental framework. Logistic Regression was selected because it provides transparent probabilistic predictions that clinicians can readily interpret. Unlike many black-box machine learning algorithms, Logistic Regression allows healthcare professionals to understand how clinical variables contribute to mortality risk, increasing clinician confidence and supporting adoption within clinical decision support systems (Hosmer, Lemeshow, & Sturdivant, 2013).

To ensure reproducibility and mitigate sampling bias, each dataset was split into training and testing subsets using an 80/20 stratified split. Stratification maintained the original class distribution across the train and test sets, and the split was implemented using scikit-learn's train_test_split with random_state=42. To evaluate model performance and assess the impact of various data balancing techniques, Stratified 5-Fold Cross-Validation was applied to the training data. This approach preserves class ratios in each fold and ensures consistent comparisons across all methods. Additionally, shuffle=True and random_state=42 were specified to promote both randomness and reproducibility in fold generation.

This study evaluates ten resampling strategies: No Resampling (Baseline), Random Oversampling (ROS), SMOTE, Borderline-SMOTE, ADASYN, Random Undersampling, Tomek Links, Edited Nearest Neighbors (ENN), SMOTE + Tomek Links, and SMOTE + ENN. For each method, a modular scikit-learn pipeline was constructed consisting of three key components: (1) the resampling method (if applicable), (2) feature scaling via StandardScaler (z-score normalization), and (3) Logistic Regression (LR) classifier. This pipeline ensured consistency in training procedures and eliminated data leakage between preprocessing and modeling stages.

Each pipeline was evaluated using the following six classification metrics, calculated across the cross-validation folds: Accuracy, Precision, Recall (Sensitivity), F1 score, ROC AUC, and PR AUC (Precision-Recall Area Under the Curve).

After cross-validation, each model was retrained on the full training set and evaluated on the unseen test set. Predictions were used to generate confusion matrices, summarizing classification outcomes as True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). These matrices were visualized using heatmaps to provide an intuitive view of model strengths and weaknesses under different resampling strategies. Finally, the results from each pipeline including all performance metrics, confusion matrix data, and predicted labels were compiled into structured DataFrames for systematic

comparison. This unified evaluation framework enables a thorough and reproducible analysis of how various resampling techniques influence classifier performance across differing levels of class imbalance. The next section presents these results and discusses key trends, trade-offs, and best practices identified through empirical analysis.

3.5 Evaluation Metrics

The effectiveness of the predictive models was evaluated using multiple performance measures to provide a comprehensive assessment of their ability to support clinical decision making. Because sepsis mortality represents a relatively rare outcome, reliance on overall accuracy alone can produce misleading conclusions by favoring prediction of the majority class. Accordingly, this study evaluates model performance using Accuracy, Precision, Recall, F1-score, Receiver Operating Characteristic Area Under the Curve (ROC-AUC), and Precision-Recall Area Under the Curve (PR-AUC).

Accuracy measures the proportion of correctly classified observations across both survival and mortality cases. Although accuracy provides a useful overall assessment of classifier performance, it is less informative in highly imbalanced datasets because a model may achieve high accuracy simply by predicting the majority class. In the context of sepsis mortality prediction, a model that classifies nearly all patients as survivors may achieve a high accuracy while failing to identify patients at greatest clinical risk.

Precision represents the proportion of patients predicted to experience mortality who actually died. High precision indicates that relatively few false alarms are generated, reducing unnecessary clinical interventions and minimizing alert fatigue among healthcare providers. Excessively low precision may overwhelm clinicians with false positive alerts, reducing confidence in the decision support system and potentially decreasing its effectiveness.

Recall, also referred to as sensitivity, measures the proportion of actual mortality cases correctly identified by the model. Recall is particularly important in sepsis prediction because failing to identify a patient at high risk of death may delay clinical intervention and adversely affect patient outcomes. For this reason, improving recall is often considered a primary objective in healthcare predictive analytics, even when accompanied by some increase in false positive predictions.

The F1-score provides a balanced assessment of model performance by calculating the harmonic mean of precision and recall. This measure is especially useful when evaluating models trained on imbalanced datasets because it reflects the trade-off between correctly identifying high-risk patients and limiting unnecessary alerts. Higher F1-scores indicate a more balanced classification model capable of supporting clinical decision making while minimizing operational burden.

The Receiver Operating Characteristic Area Under the Curve (ROC-AUC) evaluates the model's ability to discriminate between survival and mortality across multiple classification thresholds. Although ROC-AUC is widely reported in healthcare prediction studies, it can overestimate classifier performance when severe class imbalance exists because changes in false positive rates have relatively little influence on the metric.

For this reason, this study also evaluates the Precision-Recall Area Under the Curve (PR-AUC). Unlike ROC-AUC, PR-AUC focuses specifically on the model's ability to identify minority-class observations and provides a more informative assessment when predicting rare clinical events. Prior research has demonstrated that PR-AUC offers a more meaningful evaluation of predictive performance for highly imbalanced datasets because it directly reflects the trade-off between identifying true mortality cases and generating false positive alerts (Saito & Rehmsmeier, 2015).

While all evaluation metrics are reported, particular emphasis is placed on Recall, Precision, F1-score, and PR-AUC, as these measures most closely align with the objectives of clinical decision support. From an organizational perspective, healthcare institutions must balance the competing priorities of identifying as many high-risk patients as possible while avoiding excessive false alarms that contribute to clinician workload and alert fatigue. Consequently, the evaluation framework used in this study emphasizes metrics that support informed operational decision making rather than overall classification accuracy alone.

ISSN: 2473-4901
v12 n6841

Because predictive models intended for clinical decision support influence patient care rather than automate medical decisions, no single evaluation metric is sufficient to determine model suitability. Instead, healthcare organizations must consider multiple performance measures simultaneously and select operating thresholds that align with institutional priorities, available clinical resources, and acceptable levels of false positive and false negative outcomes. This multidimensional evaluation framework reflects the practical realities of implementing predictive analytics within healthcare information systems.

4. Results

4.1 Baseline Performance

The baseline Logistic Regression model failed to identify patients who ultimately died from sepsis despite achieving reasonable overall discrimination. This finding illustrates a common limitation of predictive modeling using severely imbalanced healthcare datasets.

4.2 Impact of Data Balancing on Sepsis Mortality Prediction

Table 1 summarizes the performance of the Logistic Regression model following the application of several commonly used data balancing techniques. The results demonstrate that severe class imbalance substantially limits the ability of an unbalanced model to identify patients who ultimately experience mortality. Although the baseline model achieved a ROC-AUC of 0.71, it failed to identify any mortality cases (Recall = 0.00), indicating that the model overwhelmingly favored prediction of patient survival. This finding illustrates the limitations of relying on conventional classification methods when mortality represents only a small proportion of clinical observations.

Method	PR AUC	Precision	Recall	F1-Score	AUC-ROC
None	0.14	0.2	0	0	0.71
ROS	0.13	0.12	0.75	0.21	0.71
SMOTE	0.13	0.12	0.71	0.2	0.7
B-SMOTE	0.14	0.12	0.7	0.21	0.7
ADASYN	0.13	0.12	0.71	0.2	0.7
RUS	0.13	0.12	0.75	0.21	0.71
Tomek	0.14	0.2	0	0	0.71
ENN	0.14	0.26	0	0	0.71
SMOTE+TOMEK	0.13	0.12	0.71	0.2	0.7
SMOTE + ENN	0.12	0.14	0.08	0.1	0.67

Table 1. Performance of Data Balancing Methods

Resampling techniques that materially changed the effective class distribution dramatically improved the model's ability to detect high-risk patients. Random Oversampling (ROS) and Random Undersampling (RUS) produced the highest Recall (0.75), while SMOTE, Borderline-SMOTE, ADASYN, and SMOTE combined with Tomek Links produced Recall of approximately 0.70–0.71. These methods act differently, but each reduces the dominance of the majority survival class during estimation. ROS increases the weight of observed minority cases through replication; RUS reduces the number of majority observations; and SMOTE-family methods create synthetic minority observations in feature space. For Logistic Regression, these changes alter the empirical class distribution presented during fitting and can shift the estimated decision boundary so that more observations are classified as mortality risk. The resulting increase in sensitivity is therefore consistent with the statistical role of

resampling rather than evidence that one technique creates intrinsically more informative clinical variables.

The same mechanism helps explain the reduction in Precision. When the fitted boundary becomes more responsive to the minority class, additional true mortality cases are identified, but more survivors also cross the classification threshold and are labeled high risk. Across most strongly rebalanced methods, Precision fell to approximately 0.12. In an operational hospital CDSS, that pattern would correspond to a larger pool of patients flagged for review, including more false-positive alerts. The present study does not measure clinician response to those alerts; rather, it quantifies the predictive trade-off that a later implementation study would need to evaluate against workflow capacity and alert burden.

The cleaning-oriented methods behaved differently. Tomek Links and ENN alone left Recall at 0.00, indicating that removing selected overlapping or locally inconsistent observations did not sufficiently alter the class imbalance for this Logistic Regression model. The more aggressive hybrid SMOTE+ENN approach increased Recall only to 0.08 and reduced ROC-AUC to 0.67. A plausible statistical explanation is that ENN cleaning removed observations near local class boundaries after synthetic oversampling, changing the training geometry in a way that reduced minority detection in this dataset. Because the study does not trace individual synthetic or removed observations, this mechanism should be interpreted as an explanation consistent with how the algorithms operate rather than a patient-level causal finding. The result nevertheless illustrates why more complex preprocessing does not necessarily yield a more useful performance profile.

The most important comparison concerns the evaluation metrics themselves. ROC-AUC remained between 0.70 and 0.71 for most methods even while Recall moved from 0.00 to as high as 0.75. ROC-AUC summarizes ranking performance across thresholds and can therefore remain relatively stable even when the classification behavior at a selected threshold changes substantially. Recall, by contrast, directly reveals the proportion of mortality cases detected at that operating point. PR-AUC adds a minority-class-focused view of ranking performance, while Precision indicates how many positive predictions are correct. Thus, the metrics answer different questions. Had ROC-AUC been considered alone, the balancing methods would appear to have little effect; Recall reveals a major change in mortality detection, while Precision reveals the corresponding false-positive burden. This divergence is the central empirical finding of the study and supports evaluating imbalanced clinical prediction with multiple complementary measures rather than a single summary statistic (Saito & Rehmsmeier, 2015).

Overall, these findings do not identify a universally optimal balancing technique. Instead, they show that preprocessing and metric selection expose different performance trade-offs that system designers and hospital decision makers would need to consider before deployment. The present dataset can establish how often mortality cases and false positives are identified under each approach; it cannot determine an institution-specific acceptable alert burden without workflow, staffing, threshold, and clinician-response data. Accordingly, the results should be interpreted as design-stage evidence for hospital-based predictive decision support rather than as a general implementation prescription for all healthcare organizations.

4.3 Implications for Clinical Decision Support

The findings have implications for the design and evaluation of hospital-based predictive clinical decision support. The study does not test an operational CDSS or clinician behavior. Instead, it demonstrates that upstream analytics choices can substantially change the information profile that a predictive component would supply to a downstream decision-support system. This distinction is important: technical model evaluation is not equivalent to implementation evaluation, but it constrains the alert sensitivity and false-positive burden that implementation must manage.

The baseline Logistic Regression model illustrates a common challenge in healthcare analytics. Although the model achieved a reasonable overall ROC-AUC, it failed to identify patients who ultimately experienced mortality. A clinical decision support system based on this model would provide little practical benefit because high-risk patients would not be flagged for additional clinical attention. In contrast, the application of data balancing techniques substantially increased the model's ability to identify patients at elevated risk of mortality, demonstrating that data preparation can have a significant impact on the operational usefulness of predictive models.

ISSN: 2473-4901
v12 n6841

At the same time, the results highlight the trade-off between improving patient identification and increasing the number of false positive alerts. Greater sensitivity increases the likelihood that clinicians intervene before patient deterioration occurs, but additional false positive alerts also increase clinician workload and may contribute to alert fatigue. Consequently, implementation decisions should not be based solely on maximizing predictive performance. Instead, healthcare organizations should consider available staffing, existing clinical workflows, and organizational tolerance for false alarms when selecting an appropriate operating threshold and data balancing strategy.

Model interpretability is also relevant at this design stage, but the present study does not measure clinician trust or acceptance. Logistic Regression was intentionally held constant because it provides transparent probability estimates and permits the effects of resampling to be isolated. Prior decision-support research shows that trust, ease of use, workflow integration, and perceived usefulness influence acceptance (Shibl et al., 2013), while recent CDSS design research emphasizes explainability, reliability, usability, and integration (Bayor et al., 2025). These findings therefore motivate, rather than substitute for, subsequent user-centered evaluation of the predictive component in practice.

From an information systems perspective, the study highlights a boundary between analytics development and system implementation. Data preparation and evaluation determine the characteristics of the predictive information artifact; implementation determines how that information is presented, integrated, interpreted, and acted upon. Decision-support design research has emphasized the value of reusable design knowledge rather than treating a DSS solely as an algorithmic artifact (Miah et al., 2020). In this study, the reusable lesson is that balancing and metric choices should be documented as design decisions because they materially alter sensitivity and alert burden even when ROC-AUC appears nearly unchanged.

Finally, these results underscore the importance of selecting evaluation measures that reflect clinical priorities. Traditional metrics such as overall accuracy and ROC-AUC may suggest acceptable model performance even when high-risk patients are consistently overlooked. For healthcare organizations implementing predictive analytics, metrics such as Recall and Precision-Recall AUC provide a more meaningful assessment of a model's ability to identify patients requiring immediate clinical attention. Aligning model evaluation with organizational objectives enables decision makers to select predictive models that are both statistically sound and operationally useful.

The effectiveness of a clinical decision support system depends not only on its ability to produce accurate predictions but also on its ability to identify patients who require timely intervention. Because sepsis mortality represents a relatively rare event, conventional performance measures may not adequately reflect a model's clinical usefulness.

Based on these findings, Table 2 summarizes design-stage considerations for hospitals evaluating predictive sepsis mortality components. These considerations are deliberately limited to what the retrospective analysis supports; deployment decisions require additional evidence concerning local workflow, staffing, alert thresholds, clinician acceptance, and prospective patient outcomes.

Finding	Practical Implication
Baseline model missed mortality cases	Evaluate Recall alongside Accuracy and threshold-independent metrics.
Data balancing increased Recall	Evaluate resampling strategies when severe imbalance suppresses minority-class detection.
Precision declined after balancing	Treat lower Precision as a potential increase in alert burden requiring local workflow evaluation.
ROC-AUC changed little	Use complementary metrics because ROC-AUC alone may mask threshold-level changes in minority detection.
Logistic Regression remained interpretable	Consider transparent models when explainability is a design requirement; test clinician trust separately during implementation.

Table 2. Practical Recommendations for Healthcare Organizations

5. Research Limitations, Conclusions, and Future Research

5.1 Limitations

This study should be interpreted in light of several limitations. First, the analysis was conducted using a single publicly available sepsis dataset. Although the Sepsis Survival Minimal Clinical Records dataset is large and representative of a severely imbalanced clinical prediction problem, patient populations, treatment protocols, and documentation practices vary across healthcare organizations. Consequently, model performance and the effectiveness of individual data balancing techniques may differ when applied to data collected from other institutions or healthcare systems.

Second, the study intentionally employed Logistic Regression as the predictive model to emphasize interpretability and facilitate discussion of clinical decision support implementation. While this approach aligns with the objective of developing transparent decision support systems, more complex machine learning algorithms may achieve different levels of predictive performance. The purpose of this study was not to identify the most accurate prediction algorithm but rather to examine how data balancing influences the practical usefulness of an interpretable predictive model.

Third, all balancing techniques were implemented using commonly accepted parameter settings rather than extensive hyperparameter optimization. Alternative parameter configurations or customized preprocessing pipelines may produce different results. Similarly, the study evaluated only data-level balancing approaches and did not consider algorithm-level methods such as cost-sensitive learning, class weighting, or focal loss.

Finally, this study evaluated predictive performance using retrospective data and should not be interpreted as an evaluation of an implemented CDSS. Clinician acceptance, workflow integration, usability, alert response, and patient outcomes were outside the scope of the analysis. Likewise, the observed Precision/Recall trade-offs quantify model behavior but do not establish an acceptable alert threshold for any particular hospital. Prospective validation and user-centered implementation research are necessary before operational deployment.

5.2 Future Research

Several opportunities exist to extend this research. First, future studies should evaluate the proposed approach using data collected from multiple healthcare organizations to determine the generalizability of the findings across different patient populations and clinical environments. Multi-institutional studies would also provide insight into how local practice patterns influence predictive performance and the effectiveness of various balancing techniques.

Second, future research should investigate the integration of predictive models into operational clinical decision support systems. While this study demonstrates that data balancing substantially improves mortality detection, additional work is needed to understand how clinicians interact with predictive alerts, how alert thresholds should be configured, and how predictive information can be presented to minimize alert fatigue while maximizing clinical usefulness.

Third, additional predictive algorithms should be examined within the same decision support framework. Although Logistic Regression offers important advantages in transparency and interpretability, future research should compare interpretable machine learning approaches such as Explainable Boosting Machines, decision trees, and calibrated ensemble methods to determine whether modest improvements in predictive performance justify increased model complexity.

Future investigations should also explore adaptive decision thresholds rather than relying on fixed classification thresholds. Healthcare organizations differ in staffing levels, patient acuity, and tolerance for false positive alerts. Dynamic threshold selection based on organizational priorities may provide greater operational value than optimizing a single statistical performance measure.

Finally, future work should examine the organizational and human factors associated with predictive analytics implementation. Factors such as clinician trust, workflow integration, governance, user training, and change management may ultimately have as much influence on successful clinical decision

support as improvements in predictive accuracy. Evaluating these organizational dimensions would provide a more comprehensive understanding of how predictive analytics can be effectively translated into routine clinical practice.

6. References

Ancker, J. S., Edwards, A., Nosal, S., Hauser, D., Mauer, E., & Kaushal, R. (2017). Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Medical Informatics and Decision Making, 17, 36. https://doi.org/10.1186/s12911-017-0430-8

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6(1), 20–29. https://doi.org/10.1145/1007730.1007735

Bayor, A. A., Li, J., Yang, I. A., & Varnfield, M. (2025). Designing Clinical Decision Support Systems (CDSS)—A User-Centered Lens of the Design Characteristics, Challenges, and Implications: Systematic Review. Journal of Medical Internet Research, 27, e63733. https://doi.org/10.2196/63733

Berner, E. S. (Ed.). (2007). Clinical decision support systems: Theory and practice (2nd ed.). Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology, 110, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Fernandez, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research, 61, 863–905. https://doi.org/10.1613/jair.1.11192

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In D. S. Huang, X. P. Zhang, & G. B. Huang (Eds.), Advances in Intelligent Computing (Vol. 3644, pp. 878–887). Springer. https://doi.org/10.1007/11538059_91

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118548387

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine, 24(11), 1716–1720. https://doi.org/10.1038/s41591-018-0213-5

Miah, S. J., Blake, J., & Kerr, D. (2020). Meta-design knowledge for Clinical Decision Support Systems. Australasian Journal of Information Systems, 24. https://doi.org/10.3127/ajis.v24i0.2049

Rudd, K. E., Johnson, S. C., Agesa, K. M., et al. (2020). Global, regional, and national sepsis incidence and mortality, 1990–2017: Analysis for the Global Burden of Disease Study. The Lancet, 395(10219), 200–211. https://doi.org/10.1016/S0140-6736(19)32989-7

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Salmi, M., Atif, D., Oliva, D., Abraham, A., & Ventura, S. (2024). Handling imbalanced medical datasets: Review of a decade of research. Artificial Intelligence Review, 57, Article 273. https://doi.org/10.1007/s10462-024-10884-2

Shashikumar, S. P., Stanley, M. D., Sadiq, I., Li, Q., Holder, A., Clifford, G. D., & Nemati, S. (2017). Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. Journal of Electrocardiology, 50(6), 739–743. https://doi.org/10.1016/j.jelectrocard.2017.08.013

Shibl, R., Lawley, M., & Debuse, J. (2013). Factors influencing decision support system acceptance. Decision Support Systems, 54(2), 953–961. https://doi.org/10.1016/j.dss.2012.09.018

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. npj Digital Medicine, 3, 17. https://doi.org/10.1038/s41746-020-0221-y

Tomek, I. (1976). Two modifications of CNN. IEEE Transactions on Systems, Man, and Cybernetics, 6(11), 769–772. https://doi.org/10.1109/TSMC.1976.4309452

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, SMC-2(3), 408–421. https://doi.org/10.1109/TSMC.1972.4309137