Strategic Framework for Disease Classification: Leveraging Benefits and Costs in Multiclass Analysis

SHELLYANN SOOKLAL* and PATRICK HOSEIN*, The University of the West Indies, Trinidad

Health datasets are often imbalanced, with a skew towards the healthy class, making traditional performance metrics like accuracy inadequate for evaluating minority class (disease-affected) predictions. Moreover, standard classifiers typically assume equal costs and benefits for decision outcomes, which is unsuitable for healthcare where misclassification implications vary. This paper addresses these challenges by integrating cost–benefit considerations into both the training and evaluation of classifiers through a generalized, data-driven framework for deriving benefit and cost parameters based on clinical indicators such as survival probabilities, treatment outcomes, and healthcare expenditure. We propose modifications to Naive Bayes and Logistic Regression algorithms for multiclass classification and compare them against hierarchical and ensemble cost-sensitive algorithms, including Hierarchical Cost-Sensitive Kernel Logistic Regression (H-CSKLR), SAMME.C2, BAdaCost, and NP-MC. Experiments across fetal health, vertebral column, and hepatitis C datasets show that the proposed multiclass Benefit-Based Logistic Regression achieved the highest benefit values—17.96 for fetal health, 37.19 for vertebral column, and 4.80 for HCV—outperforming H-CSKLR (15.36, -12.54, -3.44) and traditional accuracy-based Logistic Regression in most cases (19.01, 35.10, 4.60). A Latin Hypercube Sampling–based sensitivity analysis further confirmed the robustness of the proposed framework across diverse benefit–cost configurations. These results highlight the transparency, stability, and clinical relevance of the approach.

CCS Concepts: • **Computing methodologies** → **Machine learning**; **Machine learning algorithms**; **Bio-inspired approaches**; **Classification and regression trees**.

Additional Key Words and Phrases: Benefit-based analytics, Cost-Sensitive learning, Health analytics, Imbalance data, Multiclass Classification, Disease diagnosis

1 Introduction

In the healthcare industry, copious data streams, from patient demographics to medical imaging records, are generated through diverse channels such as sensors, surveys, advanced healthcare systems (AHS), cameras, mobile, and online applications [2, 19]. Leveraging this data for early disease detection holds immense potential to enhance patient survival rates and overall quality of life. However, a prevalent challenge in disease detection lies in accurately discerning various diseases from the same dataset. Machine learning (ML) algorithms have demonstrated remarkable predictive capabilities and have been widely employed by researchers for disease diagnosis. For instance, [21] combined multiple ML techniques for multiclass Alzheimer's disease classification, while [2] utilized a multiclass Random Forest approach to classify Asthma severity, and [22] employed a Gaussian kernel-based approach for multiclass glaucoma progression classification.

*Both authors contributed equally to this research.

Authors' Contact Information: Shellyann Sooklal, shellyann.sooklal@gmail.com; Patrick Hosein, patrick.hosein@sta.uwi.edu, The University of the West Indies, St. Augustine, Trinidad.

ACM 2637-8051/2026/7-ART

https://doi.org/10.1145/3833381

Yet, a significant hurdle with health datasets is their inherent skewness towards the healthy class, that is, they often contain a larger proportion of healthy individuals compared to those affected by diseases. In multiclass scenarios, classes representing various diseases typically have fewer samples than the healthy class, with the severity of disease classification correlating inversely with sample size. This data imbalance skews ML algorithms towards the majority class (healthy persons), resulting in misclassification of most minority classes (disease-affected samples). The core issue stems from the assumption that misclassification costs and benefits are uniform across different classes, driving ML algorithms to prioritize high accuracy. However, in medical contexts, this assumption is impractical, as correctly identifying disease-affected individuals is crucial for timely treatment. Misclassifying a diseased person may lead to delayed treatment, diminishing quality of life, and in severe cases, death. Conversely, misclassifying a healthy individual as diseased is also clinically consequential, as this can result in unnecessary treatments, adverse drug effects, psychological distress, and increased healthcare costs. For instance, misdiagnosing a diabetic patient could delay treatment which can result in deterioration of the person's health, while diagnosing a healthy individual as diabetic may initiate unnecessary lifestyle changes such as monitoring of insulin levels and changing dietary habits. Hence, different costs and benefits should be associated with misclassification and correct classification, respectively, reflecting the severity of disease detection. In multiclass scenarios, identifying certain disease classes may carry higher severity, warranting distinct costs for misclassification and benefits for correct classification.

In addition, common performance metrics for classification, such as accuracy, precision, sensitivity, specificity, F1-score, the receiver operating characteristic (ROC) curve, and the area under the ROC curve (AUC) [5, 14, 15, 43], have limitations with imbalanced data. Metrics like accuracy, precision, and F1-score depend on data distribution and are thus unsuitable for imbalanced datasets [17]. While recall and specificity are data-insensitive, they lack insight into the correct classification rates for both minority and majority classes. As [17] notes, the ROC curve may overestimate classifier performance without providing confidence intervals or statistical significance, and [15] argues that AUC/ROC can sometimes yield incoherent results. The primary drawback of these traditional metrics is that they assume equal benefit for correctly classifying each class and equal cost for misclassification. Consequently, when data skews heavily toward the majority class, these metrics can make biased classifiers appear highly effective, falsely suggesting strong overall performance. As previously mentioned, this bias can result in dangerous misclassifications: a healthy person may be incorrectly diagnosed with a disease (leading to unnecessary treatments), while a diseased individual might be overlooked, resulting in delayed intervention or worsened outcomes. As with the diabetes example, the misclassification of a diabetic patient could mean missed opportunities for early treatment, leading to complications, while misdiagnosing a healthy individual as diabetic could cause unnecessary lifestyle changes and stress.

This paper addresses the challenges of cost-sensitive multiclass disease classification by directing machine learning algorithms toward maximizing clinical benefit and minimizing diagnostic cost. Our approach integrates benefit and cost considerations directly into both the training and evaluation phases, enabling models to reflect the asymmetric consequences of different types of misclassifications. To achieve this, we extend the original benefit-based framework introduced in [38] and present a generalized methodology (Section 3.7) for deriving benefit and cost parameters that is data-driven, transparent, and applicable across diverse healthcare domains. This framework ensures that benefit and cost values are grounded in empirical evidence, such as survival probabilities, treatment outcomes, or healthcare expenditures, rather than arbitrary assumptions. We further demonstrate this through the re-derivation of benefit matrices for the fetal health dataset based on published perinatal survival data, ensuring greater clinical validity and reproducibility of parameter selection.

Building upon this foundation, we propose benefit-based extensions of multiclass Logistic Regression (BB-LR) and Naive Bayes (BB-NB) classifiers that embed benefit–cost parameters directly into their optimization processes. These models are benchmarked against conventional accuracy-based Logistic Regression, the Hierarchical Cost-Sensitive Kernel Logistic Regression (H-CSKLR) proposed by [47], and additional state-of-the-art multiclass

cost-sensitive algorithms, including SAMME.C2 [37], BAdaCost [11], and Neyman-Pearson -MC [41]. To assess robustness, a structured sensitivity analysis was performed by systematically perturbing the benefit–cost matrices through controlled scaling factors, following the principles of global sensitivity analysis [34]. Comprehensive experiments on fetal health, vertebral column, and hepatitis C datasets illustrate that the proposed benefit-based models consistently achieve superior and more stable performance across a wide range of benefit–cost configurations, demonstrating their effectiveness and reliability in clinically relevant decision-making contexts.

The rest of the paper is organized as follows. Section 2 offers a summary of prior efforts in cost-sensitive multiclass classification. In Section 3, we begin by revisiting our benefit performance metric from [38] and elucidate its adaptation for the multiclass scenario including the new generalized benefit–cost derivation framework. Subsequently, we delineate our strategies for integrating benefits and costs into multiclass Logistic Regression (LR) and Naive Bayes algorithms. Additionally, we introduce [47]'s hierarchical cost-sensitive kernel LR and outline how it can be modified with our benefit-based LR. Section 4 details the application of these algorithms to a fetal health dataset, comparing their outcomes to those of traditional accuracy-based LR. To validate and assess the robustness of our approaches, we extend our analysis to a vertebral column analysis dataset and a hepatitis C dataset, presented in Section 5. In Section 6 we discuss the interpretability and explainability of our results, as well as the importance of choosing suitable benefit and costs values. We then summarize our experiments and results and discuss future works in Section 7.

2 Related Work

Cost-sensitive multiclass classification has been applied across various domains to address challenges associated with imbalanced datasets [1, 10, 23, 26, 44, 49]. While prior studies have made significant strides in this area, our work seeks to address key gaps, including the lack of benefit-cost considerations in evaluating classifier performance and the limited interpretability of existing methods.

[33] proposed a cost-sensitive probability weighted ensemble learning method for multiclass classification. The ensemble model combines various classifiers such as decision trees, Naive Bayes, K-nearest neighbour and multilayer perceptron and the authors use True Positive weighting (TPweight) to perform experiments on 3TP Ensemble, 4TP Ensemble, 5TP Ensemble and 6TP Ensemble models. The authors achieve improved performance by combining weights using a weight voting method.

Furthermore, [37] proposed a cost-sensitive AdaBoost classifier, SAMME.C2, in order to solve the multiclass classification problem of predicting number of yearly claims based on telematics data on driving behaviours. SAMME.C2 is a combination of SAMME [16] (a model for multiclass classification with AdaBoost) and Ada.C2 [40] (a method for introducing costs into the AdaBoost classifier) and proved to perform better than other models such as SAMME, SMOTEBoost, SAMME with SMOTE and RUSBoost.

Similarly, [11] also proposed a cost-sensitive multiclass classifier utilizing the AdaBoost classifier. The proposed algorithm BAdaCost, combines multiple algorithms in order to obtain improved results. In order to achieve this, the algorithm CMEL (Cost-sensitive Multiclass Exponential Loss) was proposed which takes the loss values from multiple classifiers such as SAMME, AdaBoost, PIBoost and Cost-sensitive AdaBoost and combines them under one framework. The BAdaCost classifier proved to perform better than other multiclass cost-sensitive algorithms. However, one of the main issues with the approaches presented by [11, 33, 37] is that they rely on traditional metrics such as accuracy and F1-score, which assume equal costs and benefits across classes, making them unsuitable for healthcare scenarios where misclassification costs vary significantly.

[41] stated that the two main approaches to dealing with imbalance datasets are the cost-sensitive approach and the Neyman-Pearson (NP) approach. The authors explained that the NP approach does not require the derivation of costs values but to their knowledge it has mainly been used for binary classification. The authors proposed two algorithms which combines both NP and cost-sensitive learning in order to solve the NP multiclass (NPMC)

classification problem. The first algorithm NPMC-CX, which is based on convex optimization, showed strong consistency and performed well with parametric models but it performed poorly with non-parametric models. On the other hand, the second algorithm, NPMC-ER which employed empirical error rates, performed well with non-parametric models but required the dataset to be of a certain size.

On the other hand, [20] presented a 3 way decision theory rough set model for multiclass classification. The author explained that in order to minimize the cost of a Bayesian decision function, the model must consider acceptance, rejection and deferment. However, typical cost matrices cannot include values for all 3 categories, hence the authors proposed a method for incorporating all 3 categories by performing calculations based on a given cost matrix and also the relationships between the cost functions and the different classes, in order to derive cost functions for these 3 categories. The authors employed a multiphase approach for the cost-sensitive classification problem. They explained that through iterative training and prediction, the majority of rejection and deferment instances can be correctly classified and a K-nearest neighbour approach can then be used on the remaining samples with low acceptance confidence. The authors achieved improved accuracy and lower costs for misclassification when compared to other cost-sensitive approaches for multiclass classification. While these methods presented by both [41] and [20] demonstrate success, they often involve complex optimization techniques or multi-phase processes, which can limit their interpretability and scalability in real-world applications.

In addition, [18] explained that previously, only statistical methods were used to predict bond ratings for financial institutions, therefore, they proposed using machine learning models for improved results. Some of the classifiers that the authors used to test their hypothesis were logistic regression, support vector machine and cost-sensitive decision tree. They performed experiments using 3 different datasets and results confirmed that machine learning models can be used as a replacement to traditional approaches.

Finally, [47] proposed a hierarchical cost-sensitive kernel logistic regression learning algorithm which they demonstrated using a face recognition scenario. Their work extends that of [48] and basically simplifies the multiclass problem into two steps. First all minority classes are grouped into one class (in-group) and cost-sensitive learning is used to classify the data into the two groups (out-group, in-group). The samples classified as in-group are then passed through a cost-blind algorithm to further separate the samples into the various minority classes, since the authors assumed equal costs for all minority classes. Despite its novel hierarchical structure, the method assumes equal costs among minority classes during the second step, which may overlook critical differences in real-world applications.

Our work distinguishes itself by focusing on a benefit-based approach that explicitly incorporates class-specific costs and benefits into both training and evaluation, thereby addressing the limitations of traditional metrics. Moreover, our proposed models prioritize simplicity and interpretability, making them practical for healthcare professionals who require transparent decision-support tools. By refining the benefit performance metric introduced in [38], our methods offer a robust framework for evaluating and improving classification outcomes in imbalanced datasets. Additionally, we compare our approach directly with [47]'s method, demonstrating how our benefit-based logistic regression (LR) can enhance their hierarchical framework by addressing its shortcomings.

Ultimately, our research not only fills gaps in the literature by introducing a benefit-oriented evaluation framework but also highlights the practical implications of cost-sensitive classification in critical domains like healthcare, where decision-making requires both precision and clarity.

3 Methodology

In this study, we employ Logistic Regression (LR) and Naive Bayes (NB) as foundational models for benefit-based multiclass classification because of their interpretability, simplicity, and suitability for cost-sensitive adaptation. Logistic Regression is particularly valuable in healthcare contexts due to its probabilistic outputs, which make model predictions transparent and interpretable—an essential feature in clinical decision-making,

where understanding the rationale behind classifications can influence patient management. Furthermore, LR's linear decision boundaries and well-defined probabilistic framework facilitate the direct incorporation of class-specific benefit and cost terms into both the training and evaluation phases. This makes it an ideal baseline for exploring how benefit–cost weighting influences classification behavior in a controlled, explainable manner.

Naive Bayes, on the other hand, provides robustness and computational efficiency, making it well-suited for smaller or imbalanced healthcare datasets. Although its independence assumption simplifies feature relationships, NB performs reliably in many clinical applications and supports the integration of benefit or cost weighting through class-conditional probability adjustments. This allows NB to explicitly model the asymmetric consequences of misclassification—such as the differing clinical impacts of false positives and false negatives—while maintaining computational tractability.

While numerous advanced cost-sensitive algorithms exist, including kernel-based and ensemble methods, the use of LR and NB in this work is intentional. These foundational models provide transparent, analytically tractable environments for testing and demonstrating the theoretical properties of the proposed benefit–cost framework before extension to more complex non-linear architectures. This design choice emphasizes interpretability and methodological clarity, aligning with the broader goal of developing decision-support tools that clinicians can understand and trust. Moreover, by later comparing these interpretable models against a kernel-based approach, the Hierarchical Cost-Sensitive Kernel Logistic Regression (H-CSKLR) [47], we demonstrate that the proposed benefit-based adaptations can achieve competitive performance while retaining transparency, a critical advantage in healthcare-oriented machine learning.

In this section, we first recap the formulations for the benefit-based performance metric and the Benefit-Based Logistic Regression algorithm originally presented in [38], and extend these formulations to the multiclass setting. We then introduce a Multiclass Cost-Based Naive Bayes model adapted for benefit–cost optimization. Subsequently, we describe the (H-CSKLR) proposed by [47] as a benchmark for comparison, followed by a proposed adaptation that integrates our Benefit-Based Logistic Regression framework to enhance performance. We also present several state-of-the-art algorithms used to benchmark the performance of our proposed methods. Then, we introduce a generalized framework for deriving benefit and cost parameters, which provides a systematic and data-driven approach that can be applied across diverse healthcare domains. Finally, we highlight the main methodological extensions and experimental enhancements introduced in this study relative to our previous work [38].

3.1 Benefit-Based Performance Metric

Let us consider a binary classification problem where N samples belong to either class 0 or class 1. Let $\vec{x}$ represent the feature vector of a given sample. The classifier will generate a continuous score $s(\vec{x})$ which will be used to place the sample in either class 0 or class 1. Assuming that the classifier produces scores that are lower for samples of class 0 than class 1, we can define a threshold t, where instances with scores $s(\vec{x}) \leq t$ are placed in class 0 and instances with scores $s(\vec{x}) > t$ are placed in class 1. Let us denote the probability density function of the scores for class 0 and class 1 instances by $f_0(s)$ and $f_1(s)$, respectively. Similarly, we can denote the cumulative distribution functions for both classes by $F_0(s)$ and $F_1(s)$.

We can also define the costs and benefits associated with each class. If we denote b_{ij} as the benefit of classifying a sample which belongs to class i as class j, then we would achieve a positive benefit ($b \geq 0$) when $i = j$ since the sample is correctly classified. On the other hand, if $i \neq j$ then we would achieve a negative benefit or cost ($b < 0$) since the sample is incorrectly classified. We can also let π_j denote the prior probability of class $j \in \{0, 1\}$. Hence, $\pi_0 F_0(t)N$ represents the the probability of a sample belonging to class 0 times the probability of the sample being correctly classified multiplied by N, the total number of samples. To simplify, $\pi_0 F_0(t)N$ is the number of samples

that are expected to be classified correctly, given a threshold t. Therefore, the overall benefit is calculated as

$$B(t) = \pi_0 F_0(t) b_{00} + \pi_0 (1 - F_0(t)) b_{01} + \pi_1 F_1(t) b_{10} + \pi_1 (1 - F_1(t)) b_{11} \tag{1}$$

given threshold t.

b_{00} and b_{11} are the only benefit values which are positive, therefore we can maximize the expected benefit when $F_0(t) = 1$ and $F_1(t) = 0$. This only occurs when there is no overlap between the two distributions. Hence, we can define an upper bound for benefit as $\pi_0 b_{00} + \pi_1 b_{11}$. If B_γ represents the expected benefit for the classifier γ then its performance metric can be defined as

$$\mu_\gamma \equiv \frac{B_\gamma}{\pi_0 b_{00} + \pi_1 b_{11}}. \tag{2}$$

One should note that if $\mu_\gamma \approx 1$ then the performance of the classifier is approximately optimal. Hence, in general, the benefit $B(t)$ should be maximized by a classifier for given benefits. We can obtain optimality by first calculating the derivative of $B(t)$ with respect to t and then by setting this calculation to zero. This result of this modification is

$$f_0(t^*)\pi_0(b_{00} - b_{01}) = f_1(t^*)\pi_1(b_{11} - b_{10}). \tag{3}$$

3.2 Benefit Objective with Logistic Regression (Binary Classification)

Although in Section 3.1 we described how varying the threshold of a classifier can optimize the expected benefit produced by the classifier, the LR classifier itself does not aim to optimize the benefit function while training. Therefore, we need to modify the Logistic Regression cost function in order to accommodate for costs and benefits. Hence, let us consider the posterior probability of the positive class as calculated by LR, that is, the logistic sigmoid of a linear function of the feature vector. For example, the probability for feature vector $\vec{x_i}$ is

$$p_i = P(y = 1|\vec{x_i}) = h_\theta(\vec{x_i}) = g(\vec{\theta}^T \vec{x_i}) \tag{4}$$

where $h_\theta(\vec{x_i})$ represents the classification outcome for $\vec{x_i}$ based on $\vec{\theta}$, the parameter vector. $g(\cdot)$ refers to the logistic sigmoid function which is defined as

$$g(z) = \frac{1}{1 + e^{-z}}. \tag{5}$$

The LR cost function

$$J(\vec{\theta}) \equiv \frac{1}{N} \sum_{i=1}^{N} J_i(\vec{\theta}) \tag{6}$$

is minimized in order to establish parameters $\vec{\theta}$. $J_i(\vec{\theta})$ is defined as

$$J_i(\vec{\theta}) = -y_i \log(h_\theta(\vec{x_i})) - (1 - y_i) \log(1 - h_\theta(\vec{x_i})) \tag{7}$$

This cost function, however, does not take into consideration different costs for different types of errors, that is, false negatives and false positives. Therefore, we instead consider this cost function

$$J^B(\vec{\theta}) \equiv \frac{1}{N} \sum_{i=1}^{N} J_i^B(\vec{\theta}) \tag{8}$$

which maximizes benefits b_{ij}. $J_i^B(\vec{\theta})$ is defined as

$$J_i^B(\vec{\theta}) = y_i[h_\theta(\vec{x_i})b_{11} + (1 - h_\theta(\vec{x_i}))b_{10}] + (1 - y_i)[h_\theta(\vec{x_i})b_{01} + (1 - h_\theta(\vec{x_i}))b_{00}]. \tag{9}$$

Equation 9 can be rewritten as

$$J_i^B(\vec{\theta}) = y_i h_\theta(\vec{x_i})(b_{11} - b_{10}) + b_{10}(1 - y_i)(1 - h_\theta(\vec{x_i}))(b_{00} - b_{01}) + b_{01} \tag{10}$$

Table 1. Benefit-Matrix for Benefit-Based Classification

	Predicted Class				
Actual Class	**0**	**1**	**2**	**...**	**K**
0	b_{00}	b_{01}	b_{02}	...	b_{0k}
1	b_{10}	b_{11}	b_{12}	...	b_{1k}
2	b_{20}	b_{21}	b_{22}	...	b_{2k}
...	...	...	...	...	...
K	b_{k0}	b_{k1}	b_{k2}	...	b_{kk}

This function will be maximized with respect to $\vec{\theta}$ therefore we can safely remove b_{10} and b_{01} since they are constants. Also, we can multiply by -1 to convert the problem into a minimization problem. The resulting function can then be divided by $(b_{11} - b_{10})$ since it will not affect the optimal $\vec{\theta}$. Once these changes are apply, we get the following function

$$J_i^B = -y_i h_\theta(\vec{x_i}) - (1 - y_i)(1 - h_\theta(\vec{x_i}))\eta \tag{11}$$

where

$$\eta \equiv \frac{b_{00} - b_{01}}{b_{11} - b_{10}}. \tag{12}$$

Equation 11 takes on a comparable form to that of the LR cost function, however, we now scale the error for instances of class 0 by the factor η. Therefore, we can use an altered version of the LR cost function which includes the scaling by η, as shown below

$$J_i(\vec{\theta}) = -y_i \log(h_\theta(\vec{x_i})) - \eta(1 - y_i) \log(1 - h_\theta(\vec{x_i})) \tag{13}$$

in order to account for benefits while training.

3.3 Proposed Multiclass Benefit-Based Logistic Regression (BB-LR)

One common approach for dealing with multiclass LR is by using the one-vs-all or one-vs-rest method where for a dataset with K classes, K binary LR classifiers are used [27, 35]. For example, if a dataset contains 3 classes, A, B and C then 3 binary LR classifiers would be created as follows:

(1) LR classifier 1: A vs {B,C}
(2) LR classifier 2: B vs {A,C}
(3) LR classifier 3: C vs {A,B}

If $\vec{x}$ represents the feature vector for a given sample, a score will be generated for each classifier, $s_i(\vec{x})$ where $i \in \{0, ..., K\}$. This score represents the probability of the sample belonging to class i. For example, for LR classifier 1 above, $s_1(\vec{x})$ represents the probability of $\vec{x}$ belonging to class A.

The scores from all K classifiers are then compared and the classifier with the maximum score is chosen in order to determine the class with the highest probability and hence the predicted class for $\vec{x}$. That is, if the classifier for class j has the highest probability, then $\vec{x}$ will be classified as class j.

In order to incorporate benefits and costs into this approach we first need to replace the traditional binary LR with our benefit-based logistic regression. Let us now consider the benefit matrix in Table 1.

All classes have associated benefits for correct classification ($b_{ij} \geq 0, i = j$) and different costs for misclassification ($b_{ij} < 0, i \neq j$), depending on the class it is misclassified as. For example, following from the scenario above, if an instance of class A is correctly classified then there would be a benefit associated with it. However, if the sample is misclassified as either B or C then there would be two different costs for the two types of misclassification errors. If misclassifying an instance of class A as class B is more severe than misclassifying it as class C, then the cost for misclassifying the sample as B would be higher than the cost of misclassifying it as C, and vice versa. Similarly, there are different costs for misclassifying instances of B and C as class A.

Hence, when applying benefits and costs to the one-vs-all method, we must combine the benefits and costs of the classes in the "all/rest" part of the classifier in order to determine values for b_{00}, b_{01} and b_{10}. These values can be calculated for all K classifiers as follows

$$b_{00} = \sum_{i=1}^{K} \pi_i b_{ii}, \text{ for } i \neq k, \tag{14}$$

$$b_{01} = \sum_{i=1}^{K} \pi_i b_{ik}, \text{ for } i \neq k, \tag{15}$$

and

$$b_{10} = \sum_{i=1}^{K} \pi_i b_{ki}, \text{ for } i \neq k \tag{16}$$

where k represents the classifier for class k and $k \in \{0, .., K\}$. Note that $b_{11} = b_{kk}$. η and B can then be calculated using Equations 12 and 1, respectively.

Note that for the overall benefit of the BB-LR classifier, Equation 1 can be adjusted to include all options from Table 1 as follows

$$B(t) = \sum_{i=1}^{K} \sum_{j=1}^{K} \pi_i F_{ij}(t) b_{ij} \tag{17}$$

3.4 Proposed Multiclass Cost-Based Naive Bayes (BB-NB)

We consider a dataset consisting of N samples with each sample having M attributes. Each sample belongs to exactly one of K classes. Let V_{ij} be used to denote the cost of predicting a category j when the true category is i. We assume categorical attributes and that continuous attributes are clustered to form categories. Suppose that we need to classify some new sample with attributes $\vec{x}$. Using Naive Bayes, the probability of this sample lying in category C_k is given by

$$p(C_k|\vec{x}) = \prod_{i=1}^{M} \frac{p(x_i|C_k)p(C_k)}{p(x_i)} \tag{18}$$

The probability $p(x_i|C_k)$ is computed based on the probability that category x_i occurs in category C_k in the given samples. We now compute the expected cost $F(C_k)$ of classifying the sample in category C_k. This is given by

$$F(C_k) = \sum_{q=1}^{K} p(C_q|\vec{x})V_{qk} \tag{19}$$

Therefore, if we want to minimize expected cost then we would have to minimize $F(C_k)$ over k. Hence, the prediction would be

$$C_{k*} = \arg\min_k F(C_k) = \arg\min_k \sum_{q=1}^{K} p(C_q|\vec{x})V_{qk} \tag{20}$$

This can be simplified to

$$C_{k*} = \arg\min_k \sum_{q=1}^{K} \left\{ V_{qk}\ p(C_q) \prod_{i=1}^{M} p(x_i|C_q) \right\} \tag{21}$$

Let us consider some special cases. Consider the case $V_{ij} = 0$ for $i = j$ and 1 otherwise which corresponds to maximizing accuracy. Consider Equation 20.

$$\begin{aligned} C_{k*} &= \arg\min_k \sum_{q=1}^{K} p(C_q|\vec{x})V_{qk} \\ &= \arg\min_k \{1 - p(C_q|\vec{x})\} \\ &= \arg\max_k \{p(C_q|\vec{x})\} \end{aligned} \tag{22}$$

Hence the category with highest probability is chosen as one would expect.

3.5 Hierarchical Cost-Sensitive Kernel Logistic Regression (H-CSKLR)

[47] presented a H-CSKLR where they tested their approach using a face recognition system. However, for generality, we will describe their approach without using this use case. Let us consider G which represents all instances belonging to the minority classes (minority group) and E which represents all instances belonging to the majority classes (majority group). The labels for both groups can then be defined as $y = D_1, D_2, ..., D_G$ for minority group and $y = H_1, H_2, ..., H_E$ for majority group. The costs of misclassifying an instance, $\vec{x}$, can be one of the following:

- C_{HD} - cost of misclassifying a majority group instance as minority group
- C_{DH} - cost of misclassifying a minority group instance as majority group
- C_{DD} - cost of misclassifying an minority group instance as the wrong minority group from G groups

The authors did not assign any cost for correctly classifying instances, therefore for full replication we will not assign any cost as well. The cost C_{DD} is equal for misclassifications between all G groups. In addition, the authors only considered one group for majority group instances ($E = 1$) therefore the total number of classes is $(G + 1)$. For a cost-sensitive scenario with $(G + 1)$ classes, the hypothesis $\phi(\vec{x})$, or Bayes decision rule, defined as

$$\phi(\vec{x}) = \begin{cases} D_v & \text{if } f_v \text{ is the maximum} \\ H & \text{if } f_h^* \text{ is the maximum} \end{cases} \tag{23}$$

can either be D_v or I with corresponding loss functions given by:

$$L(\vec{x}, D_v) = \sum_{g=1, g\neq v}^{G} P(D_g|\vec{x})C_{DD} + P(H|\vec{x})C_{HD} \tag{24}$$

and

$$L(\vec{x}, I) = \sum_{g=1}^{G} P(D_g|\vec{x})C_{DH} \tag{25}$$

where $P(D_n|\vec{x})$ and $P(H|\vec{x})$ represent $P(y = D_g|\vec{x})$ and $P(y = H|\vec{x})$, respectively, should be minimized. If the misclassification errors happen to be equal, then Equation 24 essentially becomes the case of classifying an instance based on highest posterior probability, that is, the same as traditional classification approaches.

Let us now consider the multiclass cost-sensitive kernel Logistic Regression proposed by [48]. If

$$P(D|\vec{x}) = \sum_{g=1}^{G} P(D_g|\vec{x}) \tag{26}$$

then, from Equation 24 we get

$$L(x, D_v) = P(D|\vec{x})C_{DD} - P(D_v|\vec{x})C_{DD} + P(H|\vec{x})C_{HD}. \tag{27}$$

Since $P(D|\vec{x}) + P(H|\vec{x}) = 1$, then Equation 27 becomes

$$\begin{aligned} L(x, D_v) &= (1 - P(H|\vec{x}))C_{DD} - P(D_v|\vec{x})C_{DD} + P(H|\vec{x})C_{HD} \\ &= C_{DD} + P(H|\vec{x})(C_{HD} - C_{DD}) - P(D_v|\vec{x})C_{DD}. \end{aligned} \tag{28}$$

We also get

$$L(\vec{x}, H) = \sum_{g=1}^{G} P(D_g|\vec{x})D_{DH} = P(D|\vec{x})C_{DH}. \tag{29}$$

Consequently, for $(G + 1)$ classes, the objective functions become

$$\begin{cases} C_{DD} + P(H|\vec{x})(C_{HD} - C_{DD}) - P(D_1|\vec{x})C_{DD} \\ \vdots \\ C_{DD} + P(H|\vec{x})(C_{HD} - C_{DD}) - P(D_G|\vec{x})C_{DD} \\ P(D|\vec{x})C_{DH} \end{cases} \tag{30}$$

If from every option listed in Equation 30 we subtract $C_{DD} + P(H|\vec{x})(C_{HD} - C_{DD})$ and then divide by $-C_{DD}$, then we can solve the classification problem by selecting the maximum value from

$$\begin{cases} P(D_1|\vec{x}) \\ \vdots \\ P(D_G|\vec{x}) \\ \beta P(H|\vec{x}) - \Delta \end{cases} \tag{31}$$

where we define β as

$$\beta = \frac{C_{DH} + C_{HD} - C_{DD}}{C_{DD}} \tag{32}$$

and Δ as

$$\Delta = \frac{C_{DH} - C_{DD}}{C_{DD}}. \tag{33}$$

Let's now consider the cost-blind approach to multiclass logistic regression proposed by [50]. For a given sample, $\vec{x}$, the condition probability is defined as

$$P(D_v|\vec{x}) = \frac{e^{f(x_i)}}{1 + \sum_{g=1}^{G} e^{f_g(x_i)}}, \text{ for } v = 1, ..., G \tag{34}$$

and

$$P(H|\vec{x}) = \frac{1}{1 + \sum_{g=1}^{G} e^{f_g(\vec{x}_i)}}. \tag{35}$$

For each class, the decision function now becomes

$$\begin{cases} f_v = \ln \frac{P(D_v|\vec{x})}{P(H|\vec{x})}, & \text{for } f_1, ..., f_G \\ f_0 = \ln \frac{P(H|\vec{x})}{P(H|\vec{x})} = 0. \end{cases} \tag{36}$$

In order to classify an instance, the maximum value from Equation 36 is selected. In order to include cost-sensitivity, the function f_o^* is defined as

$$f_h^* = \ln \frac{\beta P(H|\vec{x}) - \Delta}{P(H|\vec{x})}. \tag{37}$$

Bayes decision rule, $\phi(\vec{x})$, can then be used to classify an instance.

It should be noted that selecting the maximum value from Equation 31 is equal to selecting the maximum value from Equation 23, therefore, for both cost-sensitive and cost-blind Logistic Regression, the training steps are the same and cost-sensitivity is achieved in the testing step using Equation 23.

However, using the observations from Equation 24 and the fact that the minority instances are analyzed in the later stages of the above algorithms, these approaches can be simplified using a hierarchical approach. In particular, a binary cost-sensitive classification algorithm can first be used to separate minority group from majority group samples. Then, since the cost of misclassifying minority group instances as other minority groups are the same, a cost-blind multiclass classification can be used on the minority group samples.

For binary cost-sensitive classification, a sample $\vec{x}$ is classified as minority group if $P(D|\vec{x}) \geq p^*$, where p^* is defined as

$$p^* = \frac{C_{HD}}{C_{HD} + C_{DH}} \tag{38}$$

3.6 Proposed Hierarchical Approach with Benefit-Based Logistic Regression (H-CSLR)

In order to merge the hierarchical approach proposed in Section 3.5 with the benefit-based Logistic Regression from Section 3.2, we can perform the following steps:

(1) Replace the cost-sensitive classification algorithm from Section 3.5, that is, Equation 38, with the benefit-based Logistic Regression from Section 3.2 (that is, using the cost-function from Equation 13) in order to classify instances as minority group or majority group.
(2) The minority group samples can then be classified using traditional cost-blind Logistic Regression.

3.7 Additional Comparative Algorithms

To ensure a comprehensive and balanced evaluation, we expanded our set of comparison models to include three state-of-the-art cost-sensitive learning algorithms identified in the related work: SAMME.C2, BAdaCost, and NP-MC. Each of these algorithms addresses multiclass cost sensitivity through distinct methodological formulations, providing complementary perspectives to our proposed benefit-based framework.

The SAMME.C2 algorithm extends AdaBoost to the multiclass setting by incorporating class-dependent cost adjustments into the boosting weight updates, enabling more nuanced penalization of costly misclassifications. BAdaCost further adapts this principle by introducing benefit-weighted updates that dynamically modify the boosting process based on both cost and benefit magnitudes, improving alignment with unequal class importance. The Neyman–Pearson Multi-Class (NP-MC) approach takes a probabilistic decision-theoretic perspective, optimizing classification thresholds to control type I and type II errors under explicit cost constraints.

Including these algorithms strengthens the fairness and breadth of our experimental comparison, as they represent widely recognized strategies for handling class imbalance and cost asymmetry in multiclass settings. Together with traditional Logistic Regression and the H-CSKLR, these methods provide a diverse and rigorous benchmarking suite against which to evaluate the proposed benefit-based models.

3.8 Generalized Framework for Deriving Benefit and Cost Parameters

To ensure that benefit and cost assignments are transparent, data-driven, and reproducible across healthcare domains, we propose a generalized framework for deriving these values. Benefit and cost parameters should ideally reflect real-world outcomes such as survival probabilities, treatment success rates, or healthcare expenditure. For each true class i and predicted class j, the benefit value b_{ij} can be estimated from published epidemiological data or clinical cost studies that capture the incremental change in quality-adjusted life years (QALYs), expected treatment costs, or complication rates.

For example, in oncology diagnostics, published cancer registry data and treatment outcome studies can inform the relative scaling of b_{ij} values. Suppose C_1 represents *Benign*, C_2 *Early-Stage Malignant*, and C_3 *Advanced Malignant*. The benefit of correct classification b_{ii} can be made proportional to the expected gain in survival or reduction in morbidity associated with accurate and timely diagnosis at that stage. Conversely, misclassification costs ($b_{ij} < 0$) can be derived from clinical or economic evidence quantifying the loss in survival probability, increased treatment burden, or elevated healthcare expenditure that results from diagnostic error. For instance, misclassifying an early-stage cancer as benign may lead to delayed treatment and substantial survival loss, while misclassifying a benign lesion as malignant may incur unnecessary surgical or chemotherapeutic costs. In this manner, the benefit–cost structure explicitly encodes the asymmetric clinical consequences of diagnostic errors, thereby enhancing both interpretability and decision relevance.

When empirical data are limited, expert elicitation or Bayesian updating can be used to establish prior distributions for benefit and cost parameters [3, 7, 39]. This approach improves generalizability, allowing the framework to adapt to different clinical domains such as oncology or cardiology while maintaining methodological consistency. Future work will extend this idea by automatically learning approximate benefit–cost matrices from outcome datasets, thereby minimizing subjectivity. This aligns with recent approaches in health economics that leverage Bayesian and decision-theoretic models to quantify clinical uncertainty and outcome trade-offs [30, 32].

While our framework primarily emphasizes minimizing false negatives, since missed diagnoses directly threaten patient outcomes, it also explicitly assigns costs to false positives within the benefit matrix. In our formulation, misclassifying a healthy individual as diseased incurs a negative benefit value, representing the tangible and intangible harms of over-diagnosis. These include unnecessary diagnostic procedures, medication side-effects, emotional distress, and increased healthcare expenditure. The literature highlights that false positives can produce real clinical harm; for example, [45] describe how false positive pulmonary embolism diagnoses from CT angiography often result in avoidable anticoagulation, prolonged hospitalization, and lasting mislabeling in the electronic health record. By embedding explicit penalties for such misclassifications, our model captures both ends of the clinical risk spectrum—reducing the chance of missed disease while constraining over-treatment and its associated harms. This balance ensures that optimization of the benefit metric reflects the full ethical and clinical trade-offs inherent in healthcare decision-making.

3.9 Summary of Methodological Improvements

To make the contributions of this work more explicit, Table 2 summarizes the key differences between the present study and our prior work [38]. This comparison highlights how the proposed framework extends previous research through the introduction of a generalized benefit–cost derivation methodology (Section 3.7), new multiclass model formulations, and a broader comparative evaluation that includes advanced cost-sensitive algorithms.

Table 2. Comparison between prior work and current study.

Aspect	Prior Work ([38]	Current Study
Benefit–Cost Definition	Dataset-specific benefit matrices based on domain-specific published parameters	Generalized framework (Section 3.7) for data-driven derivation using clinical or domain-informed parameters
Model Scope	Binary benefit-based Logistic Regression	BB-LR, BB-NB, and hierarchical extensions
Comparative Evaluation	Compared only to baseline LR and LR with SMOTE	Includes advanced cost-sensitive baselines (SAMME.C2, BAdaCost, NP-MC, H-CSKLR)
Sensitivity Analysis	Not included	Introduces perturbation-based sensitivity analysis of benefit–cost robustness

4 Application to Fetal Health Classification

In this section we consider the problem of fetal health classification where, based on a number of features extracted from Cardiotocogram exams, fetal health is classified as normal, suspect or pathological. We first determine appropriate benefit values and then apply these to the four multiclass algorithms described in Sections 3.3 to 3.6. We perform 5-fold cross validation and compare the results for all 4 classifiers to traditional LR which is based on optimizing accuracy rather than benefits. To assess how sensitive the proposed benefit-based algorithms are to changes in outcome valuations, we explore a wide range of benefit values that correspond to varying η thresholds. For each configuration, the performance of all four classifiers is evaluated and compared against the conventional accuracy-based Logistic Regression described in Section 3.

4.1 Dataset Description

The fetal health dataset, presented by [4], was used for our analysis since it was the only publicly available multiclass dataset we found which applied to our context. The dataset contains 21 features such as 'fetal movement', 'uterine contractions' and 'abnormal short term variability' extracted from 2126 records of Cardiotocogram exams. The records were classified by three expert obstetricians into the classes 'Normal', 'Suspect' and 'Pathological'. Based on their classifications, the dataset consists of 1655 cases of 'Normal' fetal health, 295 cases of 'Suspect' pathological and 176 cases of confirmed 'Pathological'.

4.2 Benefit and Cost Parameterization for Fetal Health Classification

Table 3 illustrates benefits b_{ij} for the fetal health dataset. To operationalize the proposed benefit–cost framework within the fetal health application, benefit values were assigned using a clinically guided estimation. Published perinatal outcome studies report survival rates exceeding 95% for normal fetuses, approximately 80–85% for suspect cases under appropriate monitoring, and below 70% for pathological cases requiring urgent intervention [9, 13, 46]. These empirically observed survival differentials were used to inform the proportional scaling of benefit and cost parameters, linking each model outcome to a realistic measure of clinical utility.

Table 3. Benefit-Matrix for Fetal Health Classification

	Predicted Class		
Actual Class	**Normal**	**Suspect**	**Pathological**
Normal	b_{00}	b_{01}	b_{02}
Suspect	b_{10}	b_{11}	b_{12}
Pathological	b_{20}	b_{21}	b_{22}

Table 4. Life Expectancy Benefit-Matrix for Fetal Health Classification

	Predicted Class		
Actual Class	**Normal**	**Suspect**	**Pathological**
Normal	15	-6	-10
Suspect	-10	20	-12
Pathological	-25	-15	30

Correct classification benefits (b_{ii}) were defined to represent the expected improvement in health outcomes or resource efficiency when each class is accurately identified. In this context, correctly classifying a *Pathological* case yields the highest clinical benefit, as timely diagnosis and intervention can substantially improve neonatal survival and prevent severe morbidity. The benefit for correctly classifying *Suspect* cases is moderate, reflecting the value of appropriate monitoring that mitigates potential complications. Correct classification of *Normal* cases carries the lowest relative benefit, corresponding to the successful continuation of routine care without unnecessary escalation. This scaling ensures that the benefit assignments reflect relative clinical impact rather than prevalence bias.

Misclassification penalties ($b_{ij} < 0$) were determined according to the severity and direction of diagnostic error. Under-diagnosis (e.g., predicting a pathological case as normal) was assigned a large negative value due to the potentially catastrophic consequences of delayed or missed intervention. Over-diagnosis (e.g., predicting a normal case as pathological) received a smaller penalty, representing the tangible but less severe consequences of unnecessary testing, anxiety, or temporary overtreatment. Intermediate penalties were applied to misclassifications involving the *Suspect* class, reflecting additional monitoring and moderate resource implications without major harm.

Table 4. summarizes the illustrative benefit–cost matrix used in model training and evaluation. The matrix was normalized to preserve interpretability while maintaining the proportional relationships across outcomes.

The diagonal entries represent the direct benefits of correct classification, scaled in proportion to the potential improvement in patient outcomes. The off-diagonal elements encode the asymmetric clinical costs of misclassification, with larger penalties assigned to missed pathological cases than to false alarms. This asymmetry mirrors clinical priorities in perinatal care, where under-detection of fetal distress poses substantially greater risk than over-detection.

By grounding benefit and cost assignments in empirical survival and treatment data, this parameterization strengthens both transparency and interpretability. It provides a consistent, domain-appropriate mechanism for translating real-world clinical outcomes into model-level optimization objectives, thereby enhancing the reproducibility and generalizability of cost-sensitive learning across healthcare applications.

Table 5. Benefit Values for Sensitivity Analysis - Fetal Health

b_{00}	b_{01}	b_{02}	b_{10}	b_{11}	b_{12}	b_{20}	b_{21}	b_{22}
16.1	-6.2	-11.1	-9.9	17.8	-12.9	-24.6	-17.6	32.5
12.8	-5.9	-10.6	-11.6	23.8	-10.1	-25.2	-15.0	35.8
14.9	-5.4	-8.5	-9.4	20.9	-12.4	-23.6	-17.3	24.1
13.9	-5.1	-11.2	-11.3	22.6	-10.7	-28.9	-16.2	25.8
17.9	-7.1	-8.8	-10.8	17.3	-14.1	-27.8	-16.1	30.0
12.1	-5.0	-11.6	-8.7	20.0	-11.4	-20.0	-14.3	33.6
16.9	-6.5	-9.4	-10.3	16.5	-11.6	-29.8	-12.8	30.3
15.5	-5.5	-10.1	-10.5	22.3	-13.4	-21.3	-15.3	27.8
13.6	-6.3	-8.4	-9.1	18.8	-13.6	-22.3	-13.5	27.5
16.7	-6.9	-9.9	-8.1	19.9	-10.4	-26.2	-12.2	31.5

4.3 Sensitivity Analysis

To further evaluate the robustness and stability of the proposed benefit-based algorithms under variations in benefit–cost assumptions, we conducted a structured sensitivity analysis. Instead of relying on a single benefit matrix or fixed multiplicative scales, we generated a set of perturbed matrices by applying element-wise random perturbations using Latin Hypercube Sampling (LHS) [29]. Each element of the baseline benefit–cost matrix was independently varied within ±20% of its original value, simulating diverse clinical or economic conditions and testing the sensitivity of model performance to small changes in both absolute and relative outcome valuations. This approach follows the principles of global sensitivity analysis, conceptually related to Morris-type screening and variance-based methods [8, 34], which assess how model outputs respond to systematic parameter variations.

Specifically, for each of $n_{\text{samples}} = 10$ LHS configurations, every model, including the proposed BB-LR, BB-NB, H-CSLR, H-CSKLR, and benchmark algorithms SAMME.C2, BAdaCost, and NP-MC, was retrained and evaluated using five stratified folds. Both mean accuracy and computed benefit were recorded for each configuration, ensuring comparability between cost-sensitive and conventional models under perturbed benefit–cost structures. Table 5 provides the generated benefit and cost values from the sensitivity analysis.

By systematically perturbing the benefit–cost matrix and observing model responses, this analysis provides a quantitative and interpretable assessment of robustness. Models exhibiting consistent benefit and accuracy across LHS samples can be considered stable with respect to subjective benefit–cost specification—an essential property for reliable and ethical deployment in healthcare decision-support systems.

4.4 Numerical Results

4.4.1 Classification Performance. In Table 6 we see average the benefit values (B), benefit performance scores (μ) and accuracy scores achieved by all 8 classifiers from Section 3 with the life expectancy benefit matrix for fetal health, across all 5 folds. Here we can see that the proposed BB-LR and the H-CSLR gave a strong performance and benefit scores when compared to more complex approaches.

Fig. 1 illustrates the benefit performance scores for all classifiers from Section 3 across all 5 folds and compares these to the benefit performance scores achieved by using traditional accuracy based LR. Similar to Table 6, we see improved performance over traditional Accuracy-Based LR with the proposed H-CSLR and for some folds. The BB-NB, however, highly prioritized the positive cases and hence produced a negative benefit score.

Table 6. Average Benefit, Performance and Accuracy Results with the Life Expectancy Benefit Values Across All 5 Folds for Fetal Health Classification

Algorithm	**Benefit,** B	**Performance,** μ	**Accuracy**
BB-LR	17.96	0.77	0.82
BB-NB	-8.83	-0.37	0.08
H-CSKLR	15.36	0.65	0.77
H-CSLR	19.08	0.81	0.86
SAMME.C2	20.28	0.86	0.89
BAdaCost	19.74	0.83	0.88
NP-MC	15.36	0.65	0.76
Traditional Accuracy-Based LR	19.01	0.80	0.85

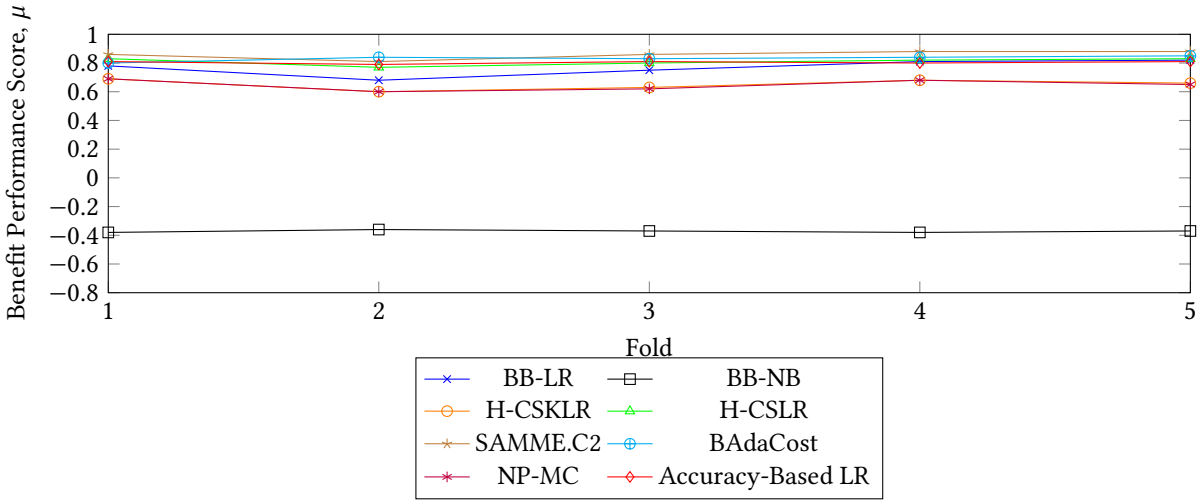

Fig. 1. Benefit Scores for Fetal Health Dataset Across 5 Folds

Ten variations in benefit matrices which resulted from the LHS were used to test the robustness of the classifiers. In Fig. 2 we plot the average benefit performance scores for all classifiers, achieved for all 10 benefit matrices across 5 folds. Here we see similar trends with the proposed algortihms to that of Fig. 1; however, there were some inconsistencies with SAMME.C2 and BAaCost in terms of which achieved the higher benefit scores.

Fig. 3 depicts the average accuracy scores for all 5 classifiers for the 10 benefit matrices across the 5 folds. We find that accuracy is sacrificed in order to achieve improved benefits.

4.4.2 Feature Importance and Model Interpretability. To further assess the interpretability of the proposed BB-LR model, coefficient-based feature importance and permutation importance analyses were performed. The coefficient estimates and corresponding odds ratios quantify the direction and magnitude of each feature's influence on the predicted class, while permutation importance evaluates the effect of each feature on the model's benefit-based

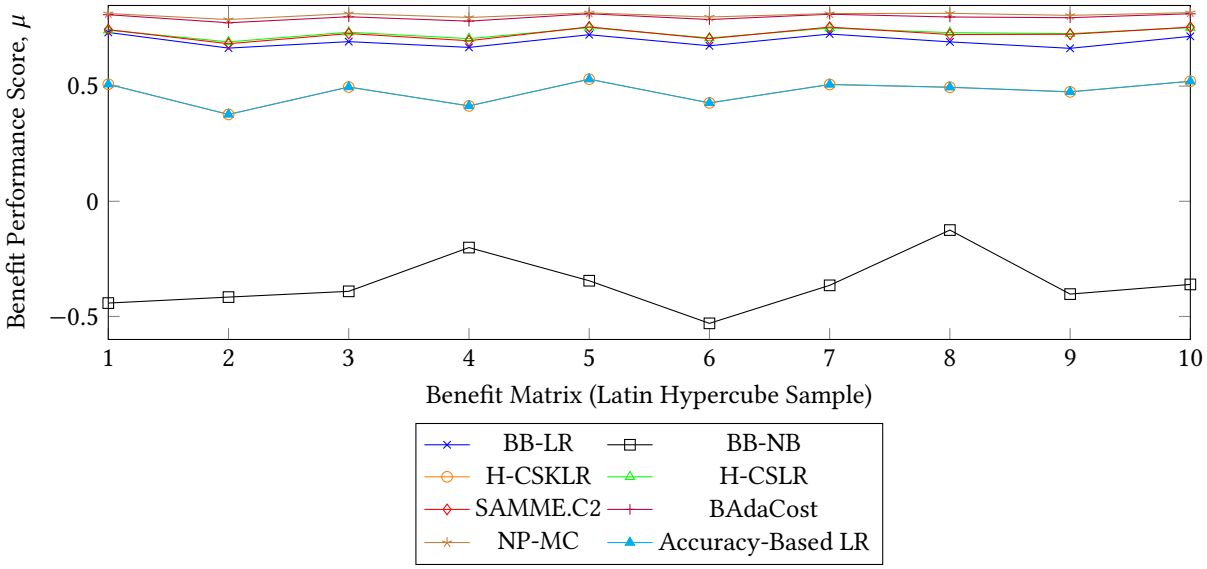

Fig. 2. Average Benefit Scores Across 10 Latin Hypercube Samples

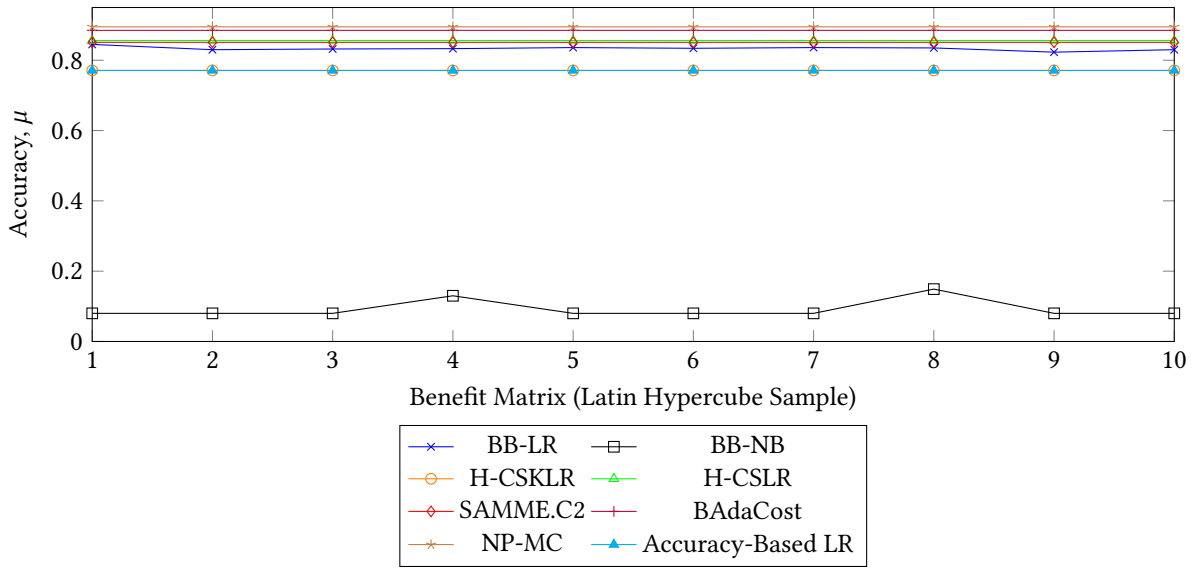

Fig. 3. Average Accuracy Across 10 Latin Hypercube Samples for All Algorithms

predictive performance. Together, these complementary analyses provide insight into the clinical variables that most strongly influence the model's decisions.

Table 7 presents the most influential features for each one-versus-rest classifier, ranked according to the absolute magnitude of their estimated coefficients. Positive coefficients indicate an increased likelihood of membership in

the corresponding class, whereas negative coefficients decrease it. The corresponding odds ratios (ORs) quantify the magnitude of these effects.

For the Fetal Health dataset, variability-related features consistently dominated the classifier-specific models. Abnormal short-term variability was the most influential predictor for both the Normal (coefficient = -0.0691, OR = 0.9333) and Pathological (coefficient = 0.0797, OR = 1.0829) classifiers, while mean value of short-term variability was the strongest predictor for the Suspect classifier (coefficient = -0.0318, OR = 0.9687). Percentage of time with abnormal long-term variability was also among the most influential features for the Normal and Pathological classifiers, whereas histogram-based characteristics contributed primarily to distinguishing the Suspect and Pathological classes.

Permutation importance further confirmed these findings. Percentage of time with abnormal long-term variability produced the largest reduction in benefit score (7.39×10^{-2}), followed by abnormal short-term variability (4.97×10^{-2}) and histogram variance (2.90×10^{-2}). The agreement between the coefficient-based rankings and permutation importance demonstrates that variability-related features consistently contributed most to the BB-LR classification decisions.

Table 7. Top five most influential features identified by the proposed BB-LR model for each one-vs-rest classifier. Features are ranked by the absolute value of the estimated coefficient. Positive coefficients increase the likelihood of the corresponding class, while negative coefficients decrease it. Odds ratios are reported to facilitate clinical interpretation.

Classifier	Feature	Coefficient	Odds Ratio
Normal	Abnormal short-term variability	-0.0691	0.9333
	Percentage of time with abnormal long-term variability	-0.0592	0.9426
	Mean value of long-term variability	-0.0452	0.9558
	Histogram number of peaks	-0.0398	0.9610
	Mean value of short-term variability	0.0298	1.0302
Suspect	Mean value of short-term variability	-0.0318	0.9687
	Histogram number of peaks	0.0312	1.0317
	Abnormal short-term variability	0.0289	1.0293
	Mean value of long-term variability	0.0268	1.0272
	Percentage of time with abnormal long-term variability	0.0231	1.0233
Pathological	Abnormal short-term variability	0.0797	1.0829
	Percentage of time with abnormal long-term variability	0.0442	1.0452
	Histogram mode	-0.0418	0.9591
	Histogram mean	-0.0406	0.9602
	Histogram variance	0.0348	1.0354

4.5 Discussion

In Section 4.4.1 above, we illustrate the performance of eight classifiers evaluated using the Fetal Health dataset and compare their outcomes to that of the traditional accuracy-based LR. Using both the life expectancy benefit matrix and the extensive sensitivity analysis performed across ten Latin Hypercube–generated benefit matrices, we assessed the robustness of each model to variations in benefit structures.

As observed previously, the H-CSKLR approach presented by [47] performed poorly across all test cases. Both its average benefit performance ($\mu_B = 0.47$) and accuracy ($\mu_A = 0.77$) were consistently below those of traditional LR ($\mu_B = 0.73$, $\mu_A = 0.85$). Inspection of its confusion matrices revealed that the model almost

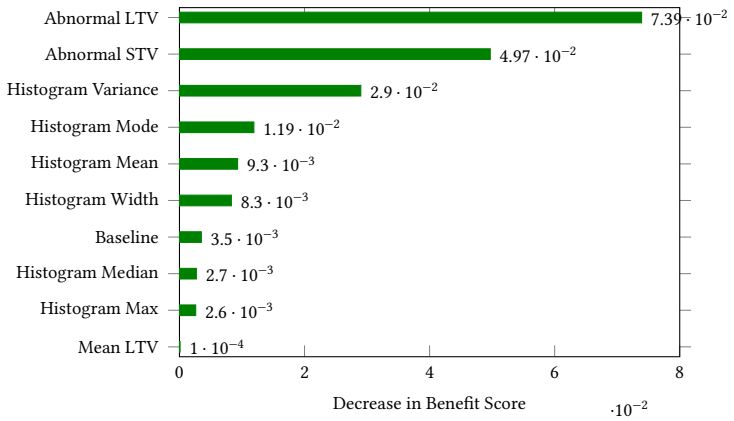

Fig. 4. Global feature importance analysis of the proposed BB-LR model using permutation importance. Features are ranked according to their contribution to the decrease in benefit score after permutation. (LTV: Long-term variability, STV: Short-term variability).

exclusively predicted the majority ("Normal") class, failing to capture the minority ("Suspected Pathological" and "Pathological") categories. This reflects the model's inability to generalize across imbalanced data distributions, a critical shortcoming in medical contexts. While optimizing for accuracy, H-CSKLR sacrifices clinical utility, an unacceptable trade-off in healthcare applications where undetected minority cases can have severe consequences. Hence, as in previous experiments, this approach was found unsuitable for fetal health classification.

Replacing the cost-sensitive component of [47]'s model with our proposed H-CSLR markedly improved performance. Across all ten benefit matrices, H-CSLR achieved high benefit scores ($\mu_B = 0.73$) and maintained accuracy at $\mu_A = 0.86$, outperforming both its cost-sensitive counterpart and traditional LR. This model consistently achieved balanced classification across all fetal health categories. Its design, which integrates benefit-driven learning within the hierarchical structure, allows it to focus on high-impact classifications without collapsing into majority-class bias. As such, H-CSLR offers the most robust trade-off between benefit optimization and diagnostic reliability, demonstrating resilience across diverse cost–benefit configurations and validating its effectiveness for healthcare applications.

The BB-NB classifier showed mixed but insightful results. It exhibited lower overall benefit ($\mu_B = -0.36$) and accuracy ($\mu_A = 0.08$–0.15) relative to other methods, reflecting its heightened sensitivity to the relative magnitudes of benefits and costs. When benefits dominated, the model favored minority classes—particularly "Suspected Pathological"—resulting in improved benefit but increased false positives. When costs were higher, the classifier reverted to majority-class bias. Although its quantitative results lag behind the others, its interpretability, probabilistic foundation, and computational efficiency make it a useful baseline for benefit-aware learning. These qualities remain advantageous in early clinical prototyping or real-time decision-support systems where model transparency and rapid inference are paramount.

The proposed BB-LR demonstrated consistently strong and stable performance across all test configurations. With an average benefit score of $\mu_B = 0.70$ and accuracy of $\mu_A = 0.83$, BB-LR outperformed traditional LR in benefit terms while maintaining acceptable accuracy loss. Its robustness to varying benefit structures underscores its ability to adaptively optimize clinical utility rather than purely statistical performance. In the confusion matrices, BB-LR achieved balanced predictions across all classes, correctly identifying high-risk fetal conditions while maintaining low false-positive rates among normal cases. This balance between clinical sensitivity and

predictive reliability highlights the model's practical value for healthcare deployment. Importantly, BB-LR achieved these gains without the heavy computational complexity or instability often associated with more sophisticated ensemble models.

When compared to the state-of-the-art ensemble methods, SAMME.C2, BAdaCost and NP-MC, notable distinctions emerge. The SAMME.C2 algorithm, which adapts AdaBoost for multiclass problems, yielded the highest mean benefit performance ($\mu_B = 0.80$) and accuracy ($\mu_A = 0.88$) across all runs. However, closer inspection revealed that its superior averages were largely due to strong performance on the majority class, with diminishing benefit gains for minority cases. Thus, while SAMME optimizes global accuracy effectively, its benefit-driven outcomes are less balanced and may not translate into better patient-level decisions.

The BAdaCost model, which explicitly integrates cost-sensitive learning into boosting, achieved the overall highest benefit score ($\mu_B = 0.82$) and accuracy ($\mu_A = 0.89$). Yet, it exhibited signs of overfitting to specific benefit configurations, its performance fluctuated when benefits and costs were close in magnitude. Although it consistently outperformed LR numerically, its stability was inferior to BB-LR and H-CSLR, indicating reduced robustness under uncertain or varying clinical benefit definitions.

Finally, the NP-MC classifier achieved identical benefit patterns to H-CSKLR ($\mu_B = 0.47$, $\mu_A = 0.77$), reaffirming that kernel-based cost formulations tend to be dominated by class imbalance unless augmented with benefit-weighted optimization. While these ensemble and kernel-based approaches achieved impressive raw scores in some cases, they lacked interpretability and adaptability compared to the proposed benefit-based models.

Overall, the results from the LHS analysis (Figures 2 and 3) confirm that the proposed BB-LR and H-CSLR achieve the most favorable balance between benefit performance, accuracy, and interpretability:

- H-CSLR demonstrates the highest clinical robustness and balance between benefit and accuracy.
- BB-LR offers a practical middle ground—strong benefit gains with stable accuracy and transparency.

While SAMME and BAdaCost achieve numerically higher accuracies, their benefit distributions and interpretability are less aligned with real-world healthcare objectives.

Furthermore, in Section 4.4.2, the interpretability analysis demonstrates that the proposed BB-LR framework provides insight into the factors influencing its classification decisions, addressing a key limitation of many machine learning approaches in healthcare, where predictions are often generated without sufficient explanation. The consistency between coefficient-based interpretation and permutation importance indicates that the identified influential predictors are robust across both class-specific and global analyses.

The predominance of variability-related features highlights the importance of temporal characteristics within cardiotocography signals for distinguishing between fetal health categories. These features capture changes in fetal heart rate behaviour and contribute to the separation of normal, suspect, and pathological conditions. By incorporating coefficient estimates, odds ratios, and feature importance measures, the proposed framework provides a transparent decision-making process that allows users to understand not only the predicted outcome but also the characteristics driving the prediction. This interpretability is particularly valuable in benefit-sensitive healthcare classification, where the consequences of different classification errors may vary. Providing insight into the factors influencing predictions can improve model transparency and facilitate greater confidence in the use of machine learning approaches for clinical decision support.

Together, these findings underscore a crucial paradigm shift: benefit-aware classifiers, particularly the proposed BB-LR and H-CSLR, move beyond accuracy as the sole performance criterion. They offer decision-support mechanisms that prioritize patient impact and clinical outcomes, aligning model optimization with the goals of modern healthcare analytics.

5 Application to Other Datasets

In order to further validate the results obtained for the fetal health dataset, we tested all 5 algorithms on two other datasets, a vertebral column classification dataset [6] and a dataset on HCV classification [25], specifically hepatitis C, fibrosis and cirrhosis.

5.1 Datasets Descriptions

The vertebral column dataset contains patient data in the form of six (6) characteristics of the pelvis and lumbar spine which were determined by their shape and alignment. These characteristics include pelvic tilt, pelvic incidence, pelvic radius, sacral slope, lumbar lordosis angle, and grade of spondylolisthesis. The goal of the dataset is to use these features to classify each patient into normal, disk hernia and spondylolisthesis. The dataset contains a total of 310 patients, of which 100 are normal, 60 are diagnosed with disk hernia and the remaining 150 are diagnosed with spondylolisthesis. This dataset is an example of a case where the majority class is not the healthy/normal patients. Therefore it is ideal for testing the robustness of the algorithms.

On the other hand, the HCV dataset comprises of laboratory measurements for both blood donors and individuals with Hepatitis C. It also contains demographic information such as sex and age. The objective of this dataset is to classify patients as normal (blood donors), hepatitis, fibrosis or cirrhosis. There are 615 patients in the dataset, 540 normal, 24 hepatitis, 21 fibrosis and 30 cirrhosis patients.

5.2 Benefit-Based Models

5.2.1 Vertebral Column. We demonstrate an approach to calculating benefits with respect to medical costs. Again, we can determine benefits starting with the baseline case. This is the case where we do nothing for the patient, that is, the person was never a patient and hence was not examined for abnormalities of the spine. If nothing is done for a 'normal' patient then there is no benefit, therefore we set $b_{00} = 0$. However, if nothing is done for a person with disk hernia then this can result in the patient eventually having to do surgery when it is finally discovered. According to [31], the cost for surgery for disk hernia can be a maximum of 50,000 USD therefore we set $b_{10} = -50$. Similarly, if nothing is done for a patient with spondylolisthesis, then they may also end up having to get surgery to correct it. According to [12], the cost of surgery for spondylolisthesis is approximately 86,000 USD, therefore we let $b_{20} = -86$.

On the other hand, if a person has disk hernia and they were correctly diagnosed, then this can lead to the person being treated and in the best case scenario they would not need to get surgery. According to [42], the cost of treatment for disk hernia is approximately 14,000 USD, therefore if we minus this from the cost of surgery to determine the amount saved, we get 36,000 USD. Therefore, we let $b_{11} = 36$. However, if a person with disk hernia is misdiagnosed as having spondylolisthesis then they will be administered the incorrect treatment. According to [36], the cost of opioid treatment for spondylolisthesis is approximately 20,000 USD. If we minus the cost of treatment for disk hernia from the cost of treatment for spondylolisthesis we get 6,000 USD. This is the excess cost that will be paid in treatment fees and also the value for b_{12}, therefore we set $b_{12} = -6$

Similarly, if a person has spondylolisthesis and they are correctly diagnosed, then this can lead to the person being treated and, again, in the best case scenario they would not need to get surgery. Therefore, we can set $b_{22} = 64$ which is the difference between the cost of surgery and the cost of treatment. However, if a person with spondylolisthesis is misdiagnosed as having disk hernia, then the wrong treatment would be administered. Since we run the risk of spondylolisthesis getting more severe which can possibly still lead to surgery. Therefore, we set $b_{21} = -30$, which is the average cost for spinal surgery according to [24], assuming that the administered treatment still had an effect on the severity of progression of the spondylolisthesis. We use this value since it is on the lower end for spondylolisthesis surgery.

Table 8. Cost-Based Benefit-Matrix for Vertebral Column Classification

	Predicted Class		
Actual Class	**Normal**	**Disk Hernia**	**Spondylolisthesis**
Normal	0	-14	-20
Disk Hernia	-50	36	-6
Spondylolisthesis	-86	-30	64

Table 9. Benefit Values for Sensitivity Analysis - Vertebral Column

b_{00}	b_{01}	b_{02}	b_{10}	b_{11}	b_{12}	b_{20}	b_{21}	b_{22}
0.0	-12.4	-20.9	-46.2	35.8	-5.3	-86.5	-29.4	72.5
0.0	-14.4	-20.4	-49.0	31.1	-4.8	-72.7	-35.0	68.2
0.0	-13.6	-21.7	-44.3	31.7	-7.2	-101.4	-24.7	69.1
0.0	-11.7	-22.9	-52.2	34.2	-5.6	-84.9	-31.4	65.9
0.0	-16.5	-23.7	-41.7	42.6	-6.5	-76.9	-31.0	52.8
0.0	-12.1	-21.6	-47.1	31.0	-7.6	-83.7	-33.2	61.7
0.0	-15.0	-24.4	-54.2	38.3	-7.2	-90.1	-27.4	55.5
0.0	-18.3	-22.3	-39.1	40.3	-6.8	-75.7	-37.6	46.9
0.0	-13.8	-19.0	-49.5	38.4	-5.5	-89.9	-29.3	72.4
0.0	-12.4	-18.4	-51.9	33.6	-5.0	-74.0	-35.2	58.9

Finally if a normal person was misdiagnosed as having disk hernia or spondylolisthesis then treatment would be administered accordingly, Based on the cost of treatment of both injuries and with the assumption that the patient is correctly diagnosed after at least one month of treatment then we can set $b_{01} = -14$ and $b_{02} = -20$. Table 8 depicts these benefit values.

5.2.2 Hepatitis C. Benefit values can be derived using life expectancy or medical costs as depicted for the fetal health and vertebral column datasets, respectively; however, here we assign values based on severity of the disease. According to [28], hepatitis C progresses to liver fibrosis and the last stage of progression is cirrhosis. Using this information benefit values are assigned as shown in Table 10.

5.3 Sensitivity Analysis

As with the fetal health dataset, sensitivity analysis was performed to test the robustness of the models with the vertebral column and HCV datasets. This analysis employed the LHS-based perturbation strategy, which systematically varies the benefit–cost parameters across a multidimensional space to generate diverse yet evenly distributed combinations of η values. Hence, the 10 combinations generated through LHS sampling test different extremities of the η parameter space. These variations, which ensure a comprehensive exploration of the model's sensitivity to benefit–cost assumptions, are summarized in Tables 9 and 11.

Table 10. Severity-Based Benefit-Matrix for HCV Classification (Normal, Hepatitis, Fibrosis, Cirrhosis)

	Predicted Class			
Actual Class	**Normal**	**Hepatitis**	**Fibrosis**	**Cirrhosis**
Normal	0	-1	-2	-3
Hepatitis	-15	25	-5	-6
Fibrosis	-32	-20	50	-15
Cirrhosis	-50	-40	-30	80

Table 11. Benefit Values for Sensitivity Analysis - HCV

b_{00}	b_{01}	b_{02}	b_{03}	b_{10}	b_{11}	b_{12}	b_{13}	b_{20}	b_{21}	b_{22}	b_{23}	b_{30}	b_{31}	b_{32}	b_{33}
0.0	-0.9	-2.2	-2.7	-17.6	25.9	-5.7	-5.2	-30.3	-22.1	41.3	-14.4	-49.7	-32.5	-27.9	75.2
0.0	-0.9	-1.9	-3.5	-16.3	28.6	-4.8	-5.9	-32.5	-16.1	58.6	-12.6	-57.1	-46.5	-34.5	95.3
0.0	-1.1	-1.8	-3.3	-16.1	29.6	-4.3	-5.4	-28.5	-20.3	49.5	-16.0	-58.0	-36.4	-33.3	89.3
0.0	-1.0	-1.9	-3.2	-14.1	26.3	-5.4	-6.2	-38.3	-23.5	46.9	-12.6	-53.1	-45.1	-27.2	91.7
0.0	-1.1	-2.4	-2.6	-13.7	20.6	-5.0	-7.1	-36.7	-22.5	52.4	-13.5	-46.3	-43.0	-24.2	67.8
0.0	-1.0	-2.3	-2.5	-15.5	27.6	-4.5	-5.7	-33.9	-18.9	42.3	-16.5	-45.6	-40.7	-35.9	77.4
0.0	-1.1	-1.5	-2.7	-12.6	24.4	-4.0	-6.3	-31.6	-18.4	56.9	-17.3	-55.6	-39.1	-30.0	82.8
0.0	-1.0	-1.7	-2.9	-16.9	23.2	-5.5	-6.9	-26.2	-21.5	44.7	-14.9	-42.5	-35.1	-31.1	84.6
0.0	-1.0	-1.9	-3.2	-14.1	26.3	-5.4	-6.2	-38.3	-23.5	46.9	-12.6	-53.1	-45.1	-27.2	91.7
0.0	-1.2	-2.1	-3.1	-14.7	22.5	-5.8	-6.6	-27.2	-19.2	54.3	-15.1	-40.6	-43.8	-25.8	66.6

5.4 Results

5.4.1 Classification Results. In Tables 12 and 13 we see the average benefit values (B), benefit performance scores (μ) and accuracy scores achieved by all 8 classifiers from Section 3 with the medical costs benefit matrix for vertebral column classification and severity benefit matrix for HCV classification, across all 5 folds. We see that the proposed BB-LR and also the H-CSLR gave the best performance and benefit scores for the vertebral column classification. Accuracy was not sacrificed due to the properties of the dataset (distribution of classes), instead benefit was increased by correctly classifying more of the more costly cases. However, since the HCV dataset contained the typical distribution of classes (majority was normal patients), accuracy was sacrificed to achieved higher benefits. The SAMME.C2 overall performance was the highest for the HCV classification; however, it's performance for the individual folds was inconsistent. SAMME.C2 and BAdaCost gave identical performance for the vertebral column classification, illustrating the boosting algorithm's limitations with the distribution of the classes in the dataset.

Using 10 benefit matrices generated via LHS, we evaluated the robustness of all eight classifiers across the Vertebral Column and HCV datasets. Figures 5 and 7 show that the proposed BB-LR and H-CSLR achieved consistently strong and stable benefit performance, with average scores between 0.72 and 0.76 on the HCV dataset and approximately 0.90 on the Vertebral dataset. Both models outperformed the traditional accuracy-based LR

Table 12. Average Benefit, Performance and Accuracy Results with the Medical Costs Benefit Values Across All 5 Folds for Vertebral Column Classification

Algorithm	**Benefit,** B	**Performance,** μ	**Accuracy**
BB-LR	37.19	0.97	0.98
BB-NB	23.90	0.63	0.49
H-CSKLR	12.54	-0.33	0.19
H-CSLR	36.38	0.97	0.97
SAMME.C2	32.70	0.85	0.91
BAdaCost	32.70	0.85	0.91
NP-MC	36.63	0.96	0.98
Traditional Accuracy-Based LR	35.10	0.92	0.97

Table 13. Average Benefit, Performance and Accuracy Results with the Severity Benefit Values Across All 5 Folds for HCV Classification

Algorithm	**Benefit,** B	**Performance,** μ	**Accuracy**
BB-LR	4.80	0.66	0.92
BB-NB	-2.06	-0.38	0.03
H-CSKLR	-3.44	-0.63	0.90
H-CSLR	3.56	0.65	0.56
SAMME.C2	5.46	0.82	0.98
BAdaCost	2.87	0.44	0.96
NP-MC	1.78	0.31	0.87
Traditional Accuracy-Based LR	4.60	0.63	0.95

and the cost-sensitive baselines, demonstrating the effectiveness of directly incorporating benefit information into model optimization.

Although SAMME and BAdaCost recorded the highest overall average benefit values (0.83–0.85), their performance across folds was highly variable, with some runs producing very low benefit scores and others extremely high ones. This inconsistency inflated their averages, suggesting that their benefit gains were incidental to their high overall accuracy rather than reflective of consistent benefit-oriented learning. In contrast, the proposed BB-LR and H-CSLR models maintained stable benefit scores across all folds and benefit matrix configurations, confirming their robustness to varying benefit–cost structures. The BB-NB and H-CSKLR models performed poorly throughout, highlighting their limited ability to adapt to benefit-weighted objectives.

The corresponding accuracy plots (Figures 6 and 8) show that both BB-LR and H-CSLR achieved accuracies around 0.85–0.86 for the HCV dataset—slightly below ensemble methods but substantially higher than cost-based models. For the Vertebral Column dataset, both benefit-based models maintained high benefit scores while showing a marginal increase in accuracy compared to standard LR. These results confirm that the proposed

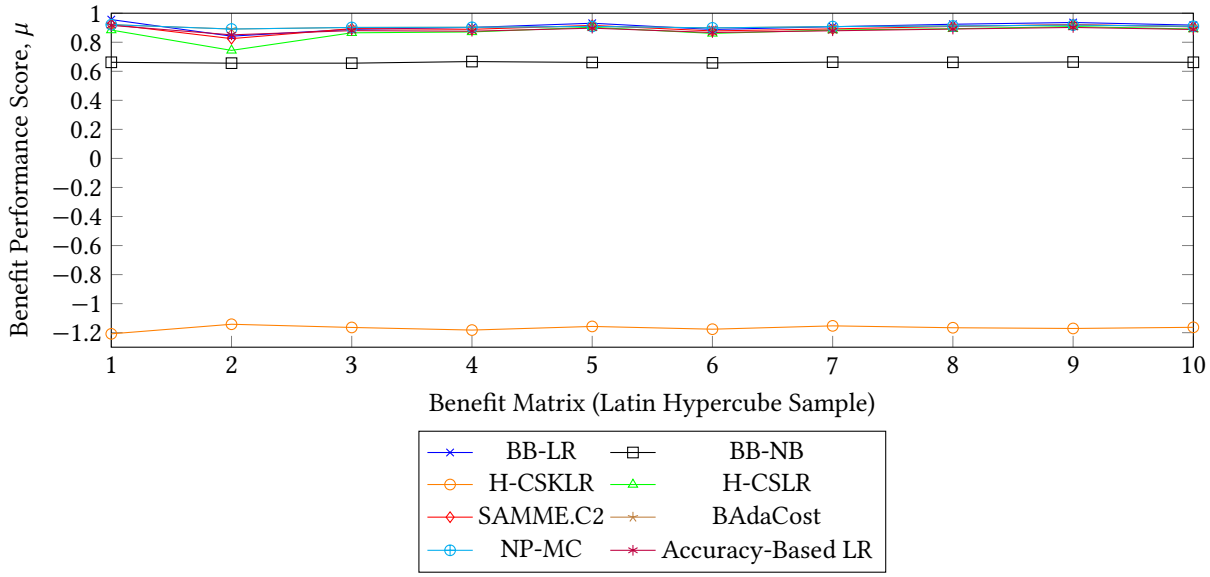

Fig. 5. Benefit Scores for Vertebral Column Dataset Across Multiple Benefit Matrices

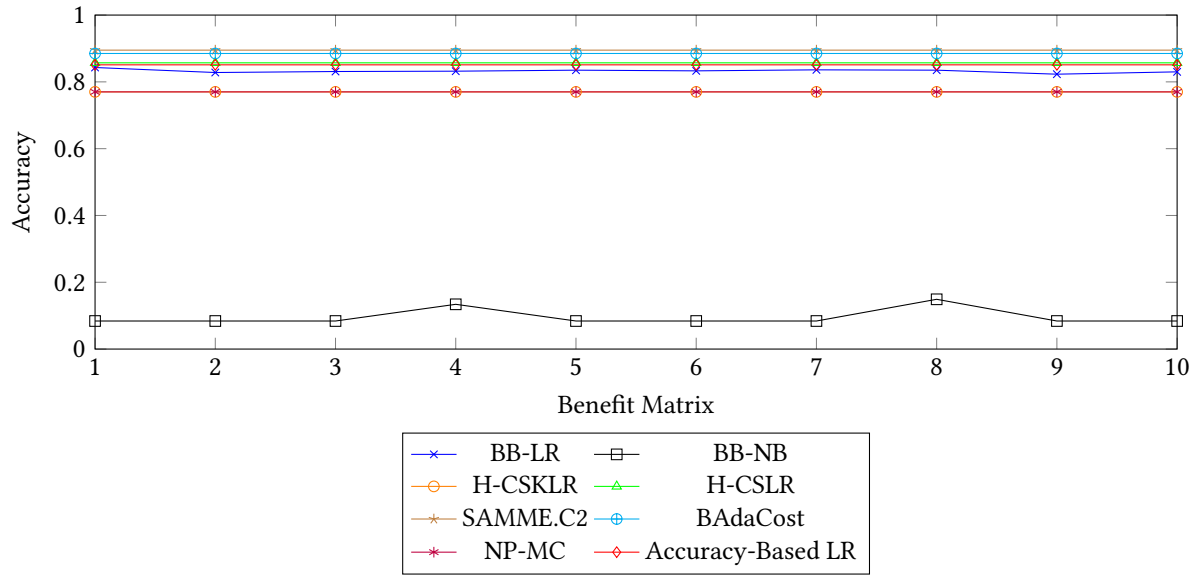

Fig. 6. Accuracy Scores for Vertebral Column Dataset Across Multiple Benefit Matrices

methods yield robust and benefit-aware classifiers capable of maintaining reliable predictive performance while optimizing for outcomes that are more meaningful in healthcare contexts.

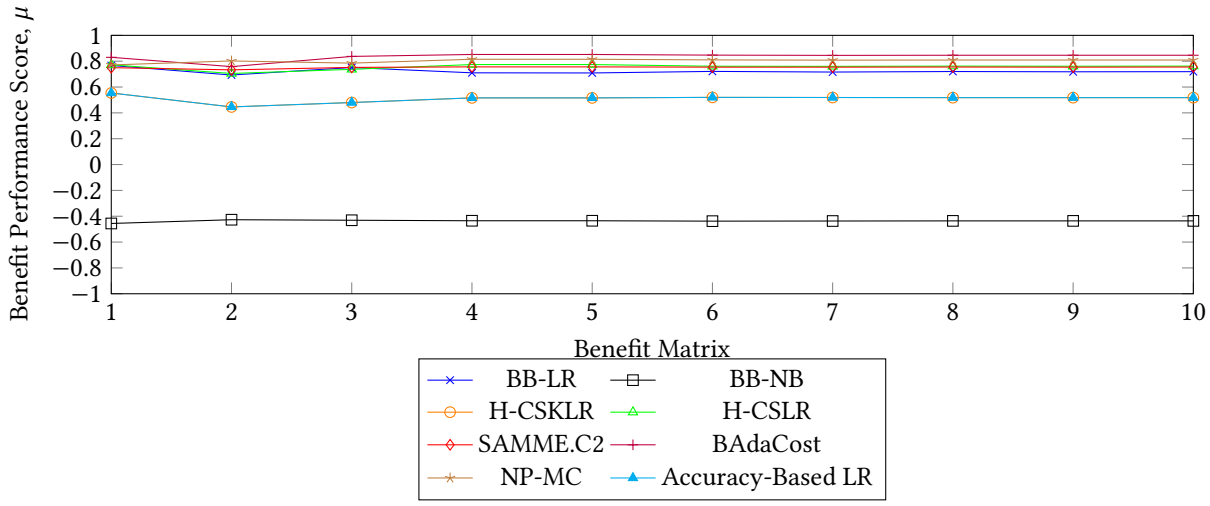

Fig. 7. Benefit Scores for HCV Dataset Across Multiple Benefit Matrices

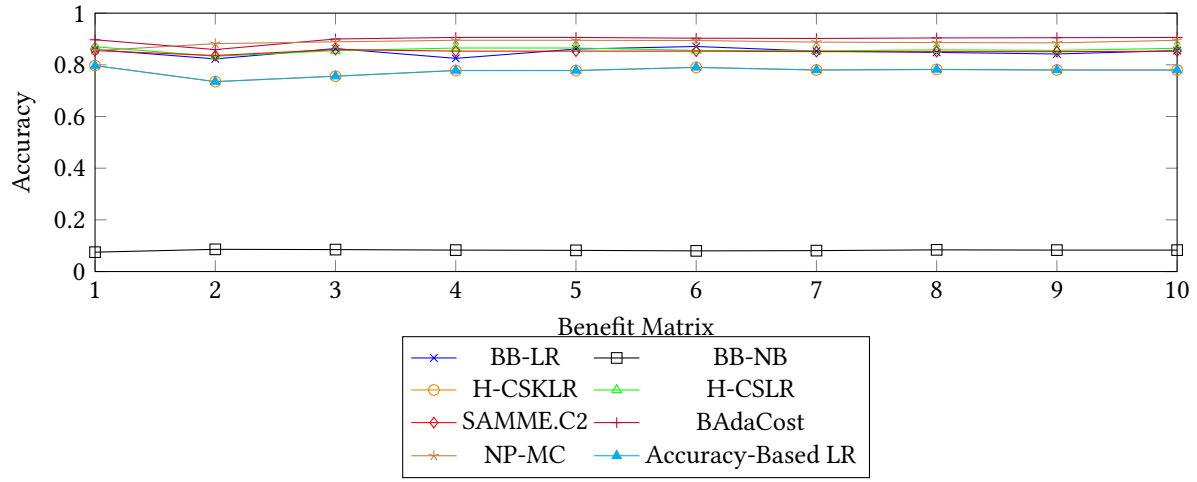

Fig. 8. Accuracy Scores for HCV Dataset Across Multiple Benefit Matrices

5.4.2 Feature Importance and Model Interpretability. Consistent with the Fetal Health dataset, model interpretability was assessed using both the estimated coefficients of the proposed BB-LR model and permutation importance. Tables 14 and 15 present the most influential features for each one-versus-rest classifier, ranked by the absolute magnitude of their coefficients. For the Vertebral Column dataset, degree spondylolisthesis was the dominant predictor across all classifiers. It exhibited the largest negative coefficient for the Normal classifier (coefficient = -0.1297, OR = 0.8784) and the largest positive coefficient for the Spondylolisthesis classifier (coefficient = 0.1836, OR = 1.2015). For the Disk Hernia classifier, pelvic tilt was the strongest positive predictor (coefficient = 0.0989, OR = 1.1040), while degree spondylolisthesis and sacral slope were negatively associated with class membership.

For the Hepatitis C dataset, the influential features varied across classifiers. ALP was the strongest positive predictor for the Normal classifier (coefficient = 0.0571, OR = 1.0587), whereas Age (coefficient = -0.0840, OR = 0.9194) and ALP (coefficient = -0.0702, OR = 0.9322) were most influential for the Hepatitis classifier. ALP also dominated the Fibrosis classifier (coefficient = -0.0751, OR = 0.9277), while albumin (ALB) had the largest effect for the Cirrhosis classifier (coefficient = -0.0424, OR = 0.9585).

Permutation importance confirmed these findings as illustrated in Figures 9 and 10. For the Vertebral Column dataset, degree spondylolisthesis produced the largest decrease in benefit score (0.4512), substantially exceeding all remaining features. For the Hepatitis C dataset, ALP (0.2110), AST (0.1962), CREA (0.1769), and ALT (0.1751) were the most influential variables, whereas Sex and CHOL had negligible importance. The consistency between the coefficient-based rankings and permutation importance demonstrates that the same predictors were identified as most influential by both local and global interpretability analyses.

Table 14. Most influential features identified by the proposed BB-LR model for the Vertebral Column dataset.

Classifier	Feature	Coefficient	Odds Ratio
Normal	Degree spondylolisthesis	-0.129655	0.878398
	Pelvic tilt	-0.096559	0.907957
	Sacral slope	0.049317	1.050554
	Pelvic radius	-0.015671	0.984451
	Pelvic incidence	-0.014521	0.985584
Disk Hernia	Pelvic tilt	0.098899	1.103955
	Degree spondylolisthesis	-0.062509	0.939404
	Sacral slope	-0.061011	0.940813
	Lumbar lordosis angle	-0.034083	0.966491
	Pelvic radius	0.015037	1.015151
Spondylolisthesis	Degree spondylolisthesis	0.183569	1.201498
	Lumbar lordosis angle	0.046844	1.047958
	Pelvic radius	-0.038194	0.962526
	Sacral slope	0.007881	1.007912
	Pelvic incidence	-0.006845	0.993178

5.5 Discussion

The results obtained from the Vertebral Column and HCV datasets, along with the sensitivity analysis using LHS, confirm the robustness and generalizability of the proposed benefit-based learning approaches. Consistent with the findings from the fetal health dataset, both the BB-LR and the H-CSLR demonstrated superior performance in terms of benefit and stability compared to the traditional accuracy-based LR and cost-sensitive baselines.

For the Vertebral Column dataset (Table 12), the proposed BB-LR model achieved the highest overall benefit (37.19) and performance score (0.97), closely followed by the H-CSLR model (36.38, 0.97). Both approaches achieved high accuracy levels (0.97–0.98), slightly exceeding the traditional accuracy-based LR (0.97) while offering higher benefit values. These results indicate that the benefit-based formulations can enhance both clinical utility and predictive precision in datasets with relatively balanced class distributions. The NP-MC method

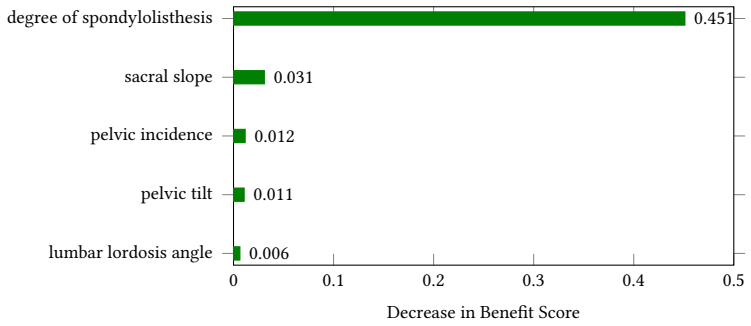

Fig. 9. Global feature importance analysis of the proposed BB-LR model using permutation importance for Vetebral Column Classification

Table 15. Most influential features identified by the proposed BB-LR model for the HCV dataset.

Classifier	Feature	Coefficient	Odds Ratio
Normal	ALP	0.057065	1.058725
	CREA	-0.049752	0.951465
	AST	-0.035603	0.965023
	GGT	-0.031709	0.968788
	PROT	-0.016420	0.983714
Hepatitis	Age	-0.084036	0.919398
	ALP	-0.070192	0.932215
	ALB	0.049044	1.050267
	PROT	0.039480	1.040270
	CHE	0.035387	1.036020
Fibrosis	ALP	-0.075059	0.927689
	ALT	0.028642	1.029057
	Age	0.019580	1.019773
	Sex	0.018968	1.019149
	GGT	0.018725	1.018902
Cirrhosis	ALB	-0.042436	0.958451
	AST	0.034447	1.035047
	BIL	0.032458	1.032991
	PROT	-0.026579	0.973771
	CHE	-0.025215	0.975100

also performed competitively (Benefit = 36.63, $\mu = 0.96$), suggesting it can adapt reasonably well when class separation is distinct. In contrast, SAMME.C2 and BAdaCost achieved good average benefit and accuracy (0.85,

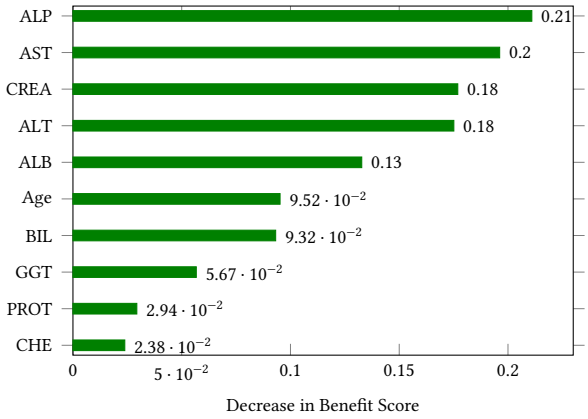

Fig. 10. Global feature importance analysis of the proposed BB-LR model using permutation importance for HCV Classification

0.91) but showed inconsistent behavior across folds, producing very high benefits in some cases and substantially lower ones in others. This variability suggests that while ensemble boosting techniques can align with benefit maximization under certain conditions, their performance lacks the fold-to-fold stability of the proposed methods. The BB-NB (Benefit = 23.90, $\mu = 0.63$) and H-CSKLR (Benefit = -12.54, $\mu = -0.33$) performed poorly, confirming that cost-based formulations fail to capture the asymmetry of benefit-driven objectives.

For the HCV dataset (Table 13), similar patterns emerged. The BB-LR and H-CSLR models achieved average benefits of 4.80 and 3.56, respectively, outperforming the traditional LR baseline (4.60). Both models maintained benefit performance scores around 0.65–0.66, demonstrating that explicit benefit optimization improves expected clinical utility even with small reductions in accuracy. The H-CSLR model exhibited slightly lower accuracy (0.56) than BB-LR (0.92), indicating greater sensitivity to class imbalance and benefit weighting. Ensemble methods again showed strong but unstable behavior. SAMME.C2 attained the highest average benefit (5.46) and accuracy (0.98), whereas BAdaCost performed inconsistently, producing benefit values ranging from high positive to near-zero across folds. Meanwhile, BB-NB and H-CSKLR continued to perform poorly, with negative benefit values and limited discriminative capability, classifying nearly all samples as dominant classes, similar to their behavior on the fetal health dataset.

The LHS sensitivity analysis across ten benefit matrices reinforces these observations. The proposed benefit-based models, particularly BB-LR, remained stable and resilient to wide variations in benefit values, consistently outperforming or matching the baseline LR in most runs. By contrast, the ensemble methods, though occasionally achieving higher peak benefits, exhibited erratic fold-level performance, while the cost-based models consistently underperformed across all benefit configurations.

In addition, the interpretability analysis demonstrated that the proposed BB-LR framework provides transparent insight into the factors influencing its predictions. The agreement between coefficient-based interpretation and permutation importance across both datasets indicates that the identified predictors are robust from both class-specific and global perspectives. For the Vertebral Column dataset, degree spondylolisthesis consistently emerged as the dominant predictor, with pelvic alignment measures also contributing to discrimination between spinal conditions. For the Hepatitis C dataset, the most influential variables corresponded to established indicators of liver function, including ALP, AST, ALT, albumin, and bilirubin, while variables such as Sex and CHOL contributed minimally.

These results solidifies the previous statements that combining model coefficients, odds ratios, and permutation importance provides a transparent and consistent explanation of the model's predictions while preserving the inherent simplicity and interpretability of logistic regression. Such interpretability is particularly valuable in benefit-sensitive healthcare applications, where understanding the rationale underlying model predictions can improve transparency and support clinical decision-making.

Overall, these findings confirm that the proposed benefit-based formulations (BB-LR and H-CSLR) provide a more robust and interpretable framework for medical classification tasks. They effectively balance benefit and accuracy, maintaining consistent performance under varying benefit structures. The stability of their results across datasets and sampling conditions highlights their potential as reliable decision-support tools for healthcare applications, where maximizing benefit is often more critical than optimizing accuracy alone.

6 Discussion

To ensure that the proposed models are practical and clinically meaningful, this study emphasizes both interpretability and methodological transparency. The proposed BB-LR model was deliberately chosen for its explainability and direct probabilistic formulation, which are essential for trust and clinical adoption. Logistic Regression offers clear, interpretable outputs that express predictions in terms of odds and likelihood, allowing clinicians to understand how individual features contribute to the final decision. Each model coefficient corresponds to a measurable influence of a patient attribute, such as a biomarker, physiological signal, or symptom, on the predicted outcome. This transparency enables healthcare professionals to trace the reasoning behind model recommendations and to integrate them confidently into diagnostic or triage workflows.

Complementing this, Naive Bayes (NB) was selected for its robustness on smaller and imbalanced datasets and for its computational efficiency in real-time clinical applications. Although NB assumes conditional independence among features, it performs reliably in many healthcare contexts and supports the incorporation of benefit–cost weighting through adjustments to its class-conditional likelihoods. Together, these foundational models provide a transparent and analytically tractable environment for testing the benefit–cost framework before extending it to more complex non-linear or ensemble architectures. Their inclusion underscores the importance of interpretability, reproducibility, and stability, key criteria for clinically deployable decision-support systems.

The integration of the generalized framework introduced in Section 3.8 further strengthens the practical relevance of the proposed approach. This framework provides a systematic, data-driven method for deriving benefit and cost parameters (b_{ij}) from measurable clinical indicators such as survival probabilities, treatment success rates, and healthcare expenditures. By grounding these parameters in epidemiological and economic data, rather than arbitrary numerical assumptions, the model's optimization objectives become more transparent and reproducible across healthcare domains. The same framework also accommodates expert elicitation and Bayesian updating when empirical data are limited, allowing it to adapt to different disease areas while maintaining methodological consistency.

Despite these advances, careful selection of benefit and cost values remains critical. Misestimated parameters can distort the model's decision boundaries, leading to clinically undesirable outcomes. For example, underestimating the cost of false negatives may cause the system to overlook high-risk patients, whereas overestimating false positives could result in unnecessary interventions, increased resource use, or patient anxiety. Such misalignments not only reduce model performance but may also introduce ethical or legal concerns if avoidable adverse outcomes occur. The sensitivity analyses conducted in this study provide a structured means to examine how variations in the benefit–cost matrix affect model behavior, supporting informed calibration that reflects clinical priorities and real-world trade-offs.

Overall, the combination of interpretable base classifiers and the generalized benefit–cost derivation framework advances the development of clinically grounded, transparent, and ethically aligned cost-sensitive learning models.

These enhancements strengthen the link between algorithmic decision-making and clinical value, fostering trust and supporting responsible deployment of machine-learning systems in healthcare.

7 Conclusions

In various domains, datasets often exhibit a skew towards specific classes, a common occurrence in healthcare datasets where healthy individuals usually predominate. Conventional machine learning classifiers typically assume uniform costs for misclassifying healthy and diseased individuals, as well as equal benefits for correctly classifying both groups. However, this assumption does not align with reality. Misclassifying a diseased individual incurs greater costs and correctly identifying such individuals holds higher benefits due to the potential for early treatment and improved recovery chances. This complexity escalates in multiclass datasets, where different classes may carry varying consequences for misclassification and benefits for correct classification.

In this paper, we address the issue of multiclass classification of imbalance datasets where the benefits and costs of correct and incorrect classification varies among the classes. We proposed a BB-LR classifier, a BB-NB classifier and also a modification to a H-CSKLR classifier proposed by [47] using our Benefit-Based LR from [38]. [47]'s method is one of the most recent works in the field, hence, we compare these three algorithms to the original classifier proposed by [47] and benchmark these against traditional accuracy-based LR as well as state-of-the-art methods, SAMME.C2, BAdaCost and NP-MC. Our comparison utilizes a Fetal Health classifier encompassing "Normal", "Suspected Pathological", and "Pathological" labels. We demonstrated how benefits and costs can be established using a life-expectancy model and applied these benefits and costs to all 5 classifiers. We access the performance of all 5 classifiers based on a benefit-based performance metric originally introduced in [38] and modified for the multiclass case, and we also show how these modifications to account for benefits and costs in the classifiers affect overall accuracy. The experiments were then repeated using a vertebral column classification dataset and a HCV classification dataset in order to validate the results obtained. The results proved that the proposed BB-LR was the most robust and stable solution out of all of the classifiers, including [47]'s method which is a recently published algorithm in the field. Furthermore, although we applied the methods for the use case of medical diagnosis, the proposed algorithm can be applied to any dataset where the benefits and costs associated with the classification outputs are important.

To provide a more systematic and clinically grounded foundation for benefit and cost estimation, we proposed a generalized framework for deriving benefit and cost parameters. This framework provides a data-driven methodology for estimating benefit and cost values (b_{ij}) from real-world clinical indicators such as survival probabilities, treatment success rates, and healthcare expenditures. The framework formalizes how these parameters can be derived from epidemiological data, clinical cost studies, or QALY estimates, while also allowing for expert elicitation and Bayesian updating in cases where empirical evidence is limited. This framework ensures a transparent, reproducible, and clinically meaningful approach for deriving benefit and cost parameters, strengthening both the theoretical rigor and real-world applicability of the proposed models.

Across all datasets and under multiple benefit configurations generated via LHS, the proposed BB-LR consistently produced the most stable and reliable benefit performance, achieving higher benefit values than traditional accuracy-based LR in nearly all cases. The H-CSLR model also performed strongly, providing comparable benefits while maintaining robust accuracy. Although ensemble-based methods such as SAMME.C2 and BAdaCost occasionally achieved higher average benefits, their performance across folds was highly inconsistent—producing very high benefits in some runs and substantially lower ones in others, resulting in inflated averages but limited reliability. In contrast, the proposed benefit-based methods remained stable across all LHS runs and datasets. The cost-sensitive methods, namely the BB-NB and H-CSKLR, performed poorly across all experiments, often favoring a single class and demonstrating limited adaptability to varying benefit structures.

Overall, the findings confirm that the proposed benefit-based formulations provide a robust and interpretable framework for multiclass healthcare classification. By explicitly optimizing for expected benefit rather than accuracy alone, the BB-LR and H-CSLR models yield decisions that are better aligned with clinical priorities. The addition of the generalized benefit-cost derivation framework further enhances the transparency and reproducibility of the approach, establishing a foundation that is adaptable across a wide range of healthcare and decision-support applications. Although demonstrated in a medical context, these methods are broadly applicable to any domain where outcomes have unequal costs and benefits.

For future endeavors, we aim to extend the benefit-based approach to other classifiers, broadening its utility across various applications. In addition, due to the sensitive nature of the results for healthcare decisions, we also would like to explore hybrid approaches which combine multiple benefit-based and cost-sensitive algorithms in order to further increase the benefits, accuracy and reliability of the results. Additionally, future research could examine the performance of these models across a broader range of datasets, particularly those with more diverse and complex clinical characteristics. For instance, datasets featuring high-dimensional data, like electronic health records (EHRs), which encompass varied patient histories, laboratory results, and diagnostic images, would test the models' robustness in handling multifaceted healthcare data. Additionally, applying these models to real-time diagnostic tools, such as those used for continuous patient monitoring or emergency care, would provide insights into their adaptability for time-sensitive applications. Exploring these areas would allow for further assessment of model accuracy, responsiveness, and scalability in real-world, high-stakes healthcare environments, helping validate their suitability for critical decision-making.

References

[1] Alekh Agarwal. 2013. Selective sampling algorithms for cost-sensitive multiclass prediction. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, Atlanta, Georgia, USA, 1220–1228. https://proceedings.mlr.press/v28/agarwal13.html

[2] Wasif Akbar, Wei-Ping Wu, Muhammad Faheem, Sehrish Saleem, Arslan Javed, and Muhammad Asim Saleem. 2020. Predictive Analytics Model Based on Multiclass Classification for Asthma Severity by Using Random Forest Algorithm. In *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. 1–4. https://doi.org/10.1109/ICECCE49384.2020.9179467

[3] D. Andrade and Y. Okajima. 2019. Efficient Bayes risk estimation for cost-sensitive classification. 89 (2019), 3372–3381. https://proceedings.mlr.press/v89/andrade19a.html

[4] Diogo Ayres-de Campos, João Bernardes, Antonio Garrido, Joaquim Marques-de Sá, and Luis Pereira-Leite. 2000. Sisporto 2.0: A program for automated analysis of cardiotocograms. *The Journal of Maternal-Fetal Medicine* 9, 5 (2000), 311–318. https://doi.org/10.3109/14767050009053454

[5] Alejandro Correa Bahnsen, Djamia Aouada, and Björn Ottersten. 2014. Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring. In *Proceedings of the 2014 13th International Conference on Machine Learning and Applications (ICMLA '14)*. IEEE Computer Society, Washington, DC, USA, 263–269. https://doi.org/10.1109/ICMLA.2014.48

[6] Guilherme Barreto and Ajalmar Neto. 2011. Vertebral Column. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5K89B.

[7] C. J. Cadham, K. Hemming, and R. E. Sherlock. 2021. The use of expert elicitation among computational modelling and health technology assessment studies: a systematic review. *BMC Medical Informatics and Decision Making* 21, 1 (2021), 72. https://doi.org/10.1186/s12911-021-01477-3

[8] F. Campolongo, J. Cariboni, and A. Saltelli. 2007. An Effective Screening Design for Sensitivity Analysis of Large Models. *Environmental Modelling & Software* 22, 10 (2007), 1509–1518. https://doi.org/10.1016/j.envsoft.2006.10.004

[9] National Perinatal Epidemiology Centre. 2021. *Perinatal Mortality Report 2021*. Technical Report. University College Cork. https://www.ucc.ie/en/media/research/nationalperinatalepidemiologycentre/NPECPMreport2021_digital.pdf

[10] Yu-An Chung, Hsuan-Tien Lin, and Shao-Wen Yang. 2016. Cost-aware Pre-training for Multiclass Cost-sensitive Deep Learning. arXiv:1511.09337 [cs.LG]

[11] Antonio Fernández-Baldera, José M. Buenaposada, and Luis Baumela. 2018. BAdaCost: Multi-class Boosting with Costs. *Pattern Recognition* 79 (2018), 467–479. https://doi.org/10.1016/j.patcog.2018.02.022

[12] Charla R Fischer, Ryan Cassilly, Marc Dyrszka, Yuriy Trimba, Austin Peters, Jeffrey A Goldstein, Jeffrey Spivak, and John A Bendo. 2016. Cost-Effectiveness of Lumbar Spondylolisthesis Surgery at 2-Year Follow-up. *Spine Deform* 4, 1 (2016), 48–54. https://doi.org/10.1016/j.jspd.2015.05.006

[13] J. Fox and S. Kumar. 2020. The impact of severe maternal morbidity on perinatal outcomes in high income countries: systematic review and meta-analysis. *Journal of Clinical Medicine* 9, 7 (2020), 2035. https://doi.org/10.3390/jcm9072035

[14] Franco Garrido, Wouter Verbeke, and Cristián Bravo. 2018. A Robust profit measure for binary classification model evaluation. *Expert Systems with Applications* 92 (2018), 154 – 160. https://doi.org/10.1016/j.eswa.2017.09.045

[15] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. 2017. Learning from Class-imbalanced Data. *Expert Syst. Appl.* 73, C (May 2017), 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

[16] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. 2009. Multi-class adaboost. *Statistics and its Interface* 2, 3 (2009), 349–360.

[17] H. He and E. A. Garcia. 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* 21, 9 (Sept 2009), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[18] Sami Ben Jabeur, Amir Sadaaoui, Asma Sghaier, and Riadh Aloui. 2020. Machine learning models and cost-sensitive decision trees for bond rating prediction. *Journal of the Operational Research Society* 71, 8 (2020), 1161–1179. https://doi.org/10.1080/01605682.2019.1581405 arXiv:https://doi.org/10.1080/01605682.2019.1581405

[19] V. Jackins, S. Vimal, M. Kaliappan, and Mi Young Lee. 2021. Prediction of Clinical Disease with AI-Based Multiclass Classification Using Naïve Bayes and Random Forest Classifier. In *Advances in Artificial Intelligence and Applied Cognitive Computing*, Hamid R. Arabnia, Ken Ferens, David de la Fuente, Elena B. Kozerenko, José Angel Olivas Varela, and Fernando G. Tinetti (Eds.). Springer International Publishing, Cham, 841–849.

[20] Xiuyi Jia, Weiwei Li, and Lin Shang. 2019. A multiphase cost-sensitive learning method based on the multiclass three-way decision-theoretic rough set model. *Information Sciences* 485 (2019), 248–262. https://doi.org/10.1016/j.ins.2019.01.067

[21] Carmen Jiménez-Mesa, Ignacio Alvarez Illán, Alberto Martín-Martín, Diego Castillo-Barnes, Francisco Jesus Martinez-Murcia, Javier Ramírez, and Juan M. Górriz. 2020. Optimized One vs One Approach in Multiclass Classification for Early Alzheimer's Disease and Mild Cognitive Impairment Diagnosis. *IEEE Access* 8 (2020), 96981–96993. https://doi.org/10.1109/ACCESS.2020.2997736

[22] Paul Y. Kim, Khan M. Iftekharuddin, Pinakin G. Davey, Márta Tóth, Anita Garas, Gabor Holló, and Edward A. Essock. 2013. Novel Fractal Feature-Based Multiclass Glaucoma Detection and Progression Prediction. *IEEE Journal of Biomedical and Health Informatics* 17, 2 (2013), 269–276. https://doi.org/10.1109/TITB.2012.2218661

[23] Yeonkook J. Kim, Bok Baik, and Sungzoon Cho. 2016. Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Systems with Applications* 62 (2016), 32–43. https://doi.org/10.1016/j.eswa.2016.06.016

[24] AM Lewin, M Fearnside, R Kuru, BP Jonkera, JM Naylor, M Sheridan, and IA Harris. 2021. Rates, costs, return to work and reoperation following spinal surgery in a workers' compensation cohort in New South Wales, 2010–2018: a cohort study using administrative data. *BMC Health Services Research* 21, 1 (2021), 1472–6963. https://doi.org/10.1186/s12913-021-06900-8

[25] Ralf Lichtinghagen, Frank Klawonn, and Georg Hoffmann. 2020. HCV data. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5D612.

[26] Xu-Ying Liu and Zhi-Hua Zhou. 2012. Towards Cost-Sensitive Learning for Real-World Applications. In *New Frontiers in Applied Data Mining*, Longbing Cao, Joshua Zhexue Huang, James Bailey, Yun Sing Koh, and Jun Luo (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 494–505.

[27] Machine Learning Mastery. 2021. One-vs-Rest and One-vs-One for Multi-Class Classification. https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/

[28] G W McCaughan and J George. 2004. Fibrosis progression in chronic hepatitis C virus infection. *Gut* 53, 3 (2004), 318–321. https://doi.org/10.1136/gut.2003.026393

[29] M. D. McKay, R. J. Beckman, and W. J. Conover. 1979. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* 21, 2 (1979), 239–245. https://doi.org/10.1080/00401706.1979.10489755

[30] W. V. Padula, P. West, and R. L. Snyder. 2022. Machine learning methods in health economics and outcomes research. *Health Economics* 31, 10 (2022), 2110–2131. https://doi.org/10.1002/hec.4834

[31] Discseel Procedure. 2020. Cost of Herniated Disc Surgery. https://discseel.com/cost-of-herniated-disc-surgery/

[32] D. Ramos-López and A. D. Maldonado. 2021. Cost-sensitive variable selection for multi-class imbalanced datasets using Bayesian networks. *Mathematics* 9, 2 (2021), 156. https://doi.org/10.3390/math9020156

[33] Artittayapron Rojarath and Wararat Songpan. 2021. Cost-sensitive probability for weighted voting in an ensemble model for multi-class classification problems. *Applied Intelligence* 51 (2021), 4908–4932. https://doi.org/10.1007/s10489-020-02106-3

[34] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. 2008. *Global Sensitivity Analysis: The Primer.* John Wiley & Sons, Chichester, UK.

[35] Towards Data Science. 2020. Multi-class Classification — One-vs-All & One-vs-One. https://towardsdatascience.com/multi-class-classification-one-vs-all-one-vs-one-94daed32a87b

[36] Mayur Sharma, Beatrice Ugiliweneza, Zaid Aljuboori, and Maxwell Boakye. 2018. Health care utilization and overall costs based on opioid dependence in patients undergoing surgery for degenerative spondylolisthesis. *Neurosurgical Focus FOC* 44, 5 (2018), E14. https://doi.org/10.3171/2018.2.FOCUS17764

[37] Banghee So, Jean-Philippe Boucher, and Emiliano A. Valdez. 2020. Cost-sensitive Multi-class AdaBoost for Understanding Driving Behavior with Telematics. https://doi.org/10.48550/ARXIV.2007.03100

[38] Shellyann Sooklal and Patrick Hosein. 2020. A Benefit Optimization Approach to the Evaluation of Classification Algorithms. In *Artificial Intelligence and Applied Mathematics in Engineering Problems*, D. Jude Hemanth and Utku Kose (Eds.). Springer International Publishing, Cham, 35–46.

[39] A. M. Stefan, A. O'Hagan, and G. Baio. 2022. Expert agreement in prior elicitation and its effects on Bayesian hypothesis testing. *BMC Medical Research Methodology* 22, 1 (2022), 134. https://doi.org/10.1186/s12874-022-01632-7

[40] Yanmin Sun, Mohamed S. Kamel, Andrew K.C. Wong, and Yang Wang. 2007. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition* 40, 12 (2007), 3358–3378. https://doi.org/10.1016/j.patcog.2007.04.009

[41] Ye Tian and Yang Feng. 2021. Neyman-Pearson Multi-class Classification via Cost-sensitive Learning. https://doi.org/10.48550/ARXIV.2111.04597

[42] Anna N. A. Tosteson, Jonathan S. Skinner, Tor D. Tosteson, Jon D. Lurie, Gunnar Andersson, Sig Berven, Margaret R. Grove, Brett Hanscom, and James N. Weinstein. 2008. The Cost Effectiveness of Surgical versus Non-Operative Treatment for Lumbar Disc Herniation over Two Years: Evidence from the Spine Patient Outcomes Research Trial (SPORT). *Spine (Phila Pa 1976)* 33, 19 (2008), 2108–2115. https://doi.org/10.1097/brs.0b013e318182e390

[43] Thomas Verbraken, Wouter Verbeke, and Bart Baesens. 2013. A Novel Profit Maximizing Metric for Measuring Classification Performance of Customer Churn Prediction Models. *IEEE Transactions on Knowledge and Data Engineering* 25 (2013), 961–973.

[44] Junhui Wang. 2013. Boosting the Generalized Margin in Cost-Sensitive Multiclass Classification. *Journal of Computational and Graphical Statistics* 22, 1 (2013), 178–192. https://doi.org/10.1080/10618600.2011.643151 arXiv:https://doi.org/10.1080/10618600.2011.643151

[45] Renda S. Wiener, Lisa M. Schwartz, and Steven Woloshin. 2013. When a test is too good: how CT pulmonary angiograms find pulmonary emboli that do not need to be found. *BMJ* 347 (2013), f3368. https://doi.org/10.1136/bmj.f3368

[46] World Health Organization. 2021. Maternal and Perinatal Death Surveillance and Response (MPDSR): materials to support implementation. https://www.who.int/teams/maternal-newborn-child-adolescent-health-and-ageing/maternal-health/maternal-and-perinatal-death-surveillance-and-response

[47] Huan Xu. 2021. Hierarchical Cost-Sensitive Techniques for Class Imbalance Learning. In *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)*. 604–609. https://doi.org/10.1109/ICAIBD51990.2021.9459083

[48] Yin Zhang and Zhi-Hua Zhou. 2010. Cost-Sensitive Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 10 (2010), 1758–1769. https://doi.org/10.1109/TPAMI.2009.195

[49] Siyuan Zhou and Ya Zhang. 2016. Active learning for cost-sensitive classification using logistic regression model. In *2016 IEEE International Conference on Big Data Analysis (ICBDA)*. 1–4. https://doi.org/10.1109/ICBDA.2016.7509840

[50] Ji Zhu and Trevor Hastie. 2004. Classification of gene microarrays by penalized logistic regression. *Biostatistics* 5, 3 (2004), 427–443.