Stability-Aware QUBO Feature Selection for Tabular Classification Under Repeated Nested Cross-Validation

Marco Fidel Mayta Quispe[1], Leonid Alemán Gonzales[2], Charles Ignacio Mendoza Mollocondo[3], Nayer Tumi Figueroa[4], Juan Carlos Juarez Vargas[5], Godofredo Quispe Mamani[6]
Instituto de Investigación en Inteligencia Computacional y Ciencia de Datos (IIICCD)
Universidad Nacional del Altiplano, Puno, Perú[1,2]
Universidad Nacional del Altiplano, Puno, Perú[3,4,5,6]

Abstract*—Quadratic Unconstrained Binary Optimization (QUBO) provides a principled framework for feature selection by encoding relevance–redundancy trade-offs and explicit constraints directly in a combinatorial objective. This study presents a stability-aware QUBO pipeline for tabular binary classification, evaluated on two standard benchmarks, namely Breast Cancer Wisconsin Diagnostic (569 samples, 30 features) and Pima Indians Diabetes (768 samples, 8 features; clinically invalid zeros treated as missing and imputed within folds). We study four QUBO variants spanning a base relevance–redundancy formulation, an exact-cardinality formulation enforcing a fixed budget k, a stability-regularized formulation that incorporates bootstrap uncertainty estimates of relevance and redundancy directly into the QUBO objective, and a performance-weighted relevance variant based on inner-CV univariate utility. All methods are assessed under repeated nested stratified cross-validation (5 outer folds $\times$ 3 repeats, $n = 15$ outer test evaluations), reporting AUC-ROC, AUC-PR, MCC, and Brier score with 95% confidence intervals, alongside selection stability via mean Jaccard similarity across outer-fold selected subsets. Results show that QUBO-based selection is competitive with strong classical baselines (RFECV, L1-logistic, permutation-importance ranking, and mutual information) while enabling strict budget control and transparent stability diagnostics. On the near-ceiling Breast Cancer benchmark, predictive differences are marginal and the main differentiators become subset-size control and stability; on Pima, QUBO-k remains competitive while enforcing strict cardinality constraints. These findings support QUBO as a practical framework when budgeted, interpretable, and reproducible feature selection is required, though evaluation is limited to low-dimensional tabular settings.

Keywords*—Feature selection; QUBO; simulated annealing; nested cross-validation; selection stability; Jaccard similarity; probability calibration; tabular classification

I. Introduction

Feature selection is a core component of tabular machine learning because it influences generalization, interpretability, and operational cost. In real deployments, the objective is often not only to maximize discrimination, but also to obtain compact and reproducible feature sets under explicit constraints (e.g., acquisition budgets, latency limits, or mandated subset sizes). Classical families—filters, wrappers, and embedded methods—offer useful trade-offs: filters are fast but may miss multivariate structure, wrappers can be accurate but costly and variance-prone, and embedded methods depend on model-specific inductive biases.

Quadratic Unconstrained Binary Optimization (QUBO) offers a unified way to encode feature selection as a single combinatorial objective balancing relevance and redundancy while supporting constraints through penalty terms [1]. This becomes attractive when practitioners need *explicit control* over subset size and a transparent assessment of *selection stability*. In correlated feature regimes, instability often arises because several variables carry overlapping predictive information and small resampling perturbations can swap one correlated surrogate for another; QUBO addresses this structurally through pairwise redundancy penalties that explicitly discourage co-selection of correlated features. Moreover, because feature selection itself is part of the learning pipeline, rigorous evaluation is essential to avoid optimistic bias; nested cross-validation is the standard remedy when selection and calibration/tuning are performed within training folds [3], [4].

We evaluate on two benchmark datasets representing distinct dimensionality regimes, namely `breast_cancer` ($d = 30$) and `pima_diabetes` ($d = 8$). To enable fair comparisons under a fixed feature budget, we benchmark selectors under target budgets k (top-k filters and QUBO variants), while also reporting realized cardinality for procedures that determine k internally (e.g., RFECV and L1-logistic).

Contributions: This work presents QUBO feature selection variants that encode relevance–redundancy trade-offs, explicit k-budgets, and stability regularization within a unified combinatorial objective. These variants are benchmarked against strong classical baselines using repeated nested stratified CV with confidence intervals across discrimination, classification-quality, and calibration metrics. In addition, selection reproducibility is quantified via mean pairwise Jaccard similarity across outer-fold selections, treating stability as a first-class criterion alongside predictive performance.

II. Related Work

Feature selection (FS) is commonly categorized into *filter*, *wrapper*, and *embedded* methods. Filter methods rank variables independently of the final predictor; mutual-information criteria are popular when nonlinear dependencies are expected, and mRMR is a canonical formulation balancing relevance to the target and redundancy among selected variables [5]. Wrapper methods evaluate subsets through downstream performance, often improving accuracy at higher computational

cost and sensitivity to the resampling design [10]. RFE and RFECV are widely used wrapper procedures that iteratively remove features based on model utility and are influential in biomedical pipelines [11]. Embedded methods incorporate selection within training; L1-regularization (LASSO) induces sparsity and performs implicit selection [6].

Regarding stability mechanisms, filters such as MI ranking are stable only when the top-k features have clearly separated relevance scores; in correlated regimes, near-tied features can be exchanged across resamples. Wrappers inherit instability from the downstream model and the elimination path, while embedded methods such as LASSO can produce multiple near-equivalent sparse solutions in correlated settings. Stability selection [7] addresses this by aggregating selection frequencies over subsamples, but it does not enforce hard cardinality constraints. The present QUBO formulation instead incorporates stability directly into the combinatorial objective via bootstrap uncertainty penalties, maintaining hard budget control within a single optimization step.

Beyond accuracy, *stability* is critical whenever FS supports interpretation or scientific conclusions. Stability selection formalizes robustness via subsampling and selection frequencies [7], while broader analysis show that unstable selectors can undermine reproducibility even under high predictive performance [8]. This motivates reporting stability metrics (e.g., Jaccard similarity) jointly with discrimination and calibration.

Rigorous evaluation is particularly important because selecting features on the same data used to estimate performance can bias results upward. Nested cross-validation mitigates this by isolating all selection/tuning within inner loops and evaluating only on held-out outer folds [3], [4]. Calibration can diverge from discrimination; probability quality should be assessed under the same leakage-safe protocol [12], [13].

Optimization-based FS treats subset selection as a combinatorial problem. Simulated annealing (SA) is a classic metaheuristic for large discrete spaces and provides a reproducible baseline for approximate minimization under fixed compute budgets [9]. QUBO offers a unifying representation for such objectives, enabling relevance–redundancy trade-offs and hard/soft constraints via quadratic penalties [1]. Prior QUBO-style feature selection formulations [2] have focused on relevance–redundancy encoding and cardinality control; the present work extends this line by integrating bootstrap uncertainty-based stability regularization into the QUBO objective and evaluating under a leakage-safe repeated nested cross-validation protocol that jointly reports discrimination, calibration, and subset reproducibility.

III. Methodology

A. Problem Setting

Let $\mathcal{D} = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$ be a labeled dataset with $\mathbf{x}^{(n)} \in \mathbb{R}^d$ and $y^{(n)} \in \{0, 1\}$.

Feature selection aims to identify a subset $S \subseteq \{1, \ldots, d\}$ that preserves predictive performance while controlling redundancy and improving interpretability under a constrained feature budget.

We encode selection using a binary vector $\mathbf{z} \in \{0, 1\}^m$ over a candidate pool of size $m \leq d$, where $z_i = 1$ indicates that candidate feature i is selected, and $S(\mathbf{z}) = \{i : z_i = 1\}$.

B. Datasets and Preprocessing (Leakage Control)

We use Breast Cancer Wisconsin Diagnostic (569 samples, 30 numeric features) [16], [17] and Pima Indians Diabetes (768 samples, 8 numeric features) [18]. For Pima, clinically invalid zeros in specific measurements (commonly including glucose, blood pressure, skin thickness, insulin, BMI) are treated as missing and imputed.

We perform all preprocessing *within the training portion of each outer fold* to prevent information leakage. We impute missing values using the training median, and features are standardized via z-score scaling (fit on training, applied to validation/test). Any dataset-specific corrections are applied inside the same fold-isolated preprocessing.

C. Compared Methods and Feature Budgets

To compare methods under explicit budgets, we define target cardinalities k and evaluate rank-based baselines returning exactly k variables (MI-top$k$, permutation-importance top$k$), QUBO variants matched to the same k target, and methods that determine subset size internally (L1-logistic and RFECV), for which we report realized cardinality. In our experiments, $k = 10$ for `breast_cancer` and $k = 6$ for `pima_diabetes`.

D. Candidate Pool Construction (for QUBO)

Within each outer training fold, we build a candidate pool by ranking features using mutual information computed on standardized training data, retaining the top candidates:

$$m = \min(25, d)$$

For `breast_cancer`, this yields $m = 25$ out of $d = 30$, while for `pima_diabetes` ($d = 8$) the candidate pool equals the full feature set ($m = d$). All QUBO objectives are constructed and optimized only over this candidate pool.

E. Relevance and Redundancy Signals

1) Relevance: For each candidate feature i, we compute mutual information relevance:

$$r_i = I(X_i; Y), \tag{1}$$

estimated with a nonparametric MI estimator (e.g., as implemented by `mutual_info_classif`).

2) Redundancy: To discourage selecting overlapping variables, we compute pairwise redundancy via absolute Pearson correlation on standardized training data:

$$\rho_{ij} = |\mathrm{corr}(X_i, X_j)|. \tag{2}$$

F. Base QUBO Objective

Let m denote the number of candidates. The base QUBO energy is given by:

$$\min_{\mathbf{z}\in\{0,1\}^m} E(\mathbf{z}) = \sum_{i=1}^{m}(\lambda - r_i)z_i + \alpha \sum_{i<j} \rho_{ij}\, z_i z_j, \qquad (3)$$

where, $\alpha > 0$ controls redundancy and $\lambda \geq 0$ encourages sparsity.

G. Exact-Cardinality QUBO-k

To target a subset size k, we add a quadratic penalty:

$$E_k(\mathbf{z}) = E(\mathbf{z}) + \mu \left(\sum_{i=1}^{m} z_i - k \right)^2, \qquad (4)$$

with $\mu > 0$ controlling constraint strength.

1) Deterministic exact-k projection (budget-matched evaluation): Because heuristic solvers may return $|S(\mathbf{z})| \neq k$ and because we compare methods under a fixed budget, we apply a deterministic projection step for the QUBO family. When $|S| > k$, we keep the k features in S with highest MI relevance; when $|S| < k$, we fill with remaining candidates with highest MI. This ensures budget-matched evaluation for all QUBO variants.

H. Stability-Regularized QUBO

To promote robustness, we estimate uncertainty of relevance and redundancy via bootstrap resampling of the outer training fold. For bootstrap replicates $b = 1, \ldots, B$, we compute $r_i^{(b)}$ and $\rho_{ij}^{(b)}$ and estimate means and variances as:

$$\bar{r}_i = \mathbb{E}[r_i^{(b)}], \qquad \mathrm{Var}(r_i) = \mathrm{Var}(r_i^{(b)}), \qquad (5)$$

$$\bar{\rho}_{ij} = \mathbb{E}[\rho_{ij}^{(b)}], \qquad \mathrm{Var}(\rho_{ij}) = \mathrm{Var}(\rho_{ij}^{(b)}). \qquad (6)$$

We then define a stability-regularized objective as:

$$E_s(\mathbf{z}) = \sum_{i=1}^{m} \Big(\lambda - \bar{r}_i + \gamma\, \mathrm{Var}(r_i)\Big) z_i + \sum_{i<j} \Big(\alpha\, \bar{\rho}_{ij} + \delta\, \mathrm{Var}(\rho_{ij})\Big) z_i z_j, \qquad (7)$$

where, $\gamma, \delta \geq 0$ weight robustness penalties. As with other QUBO variants, we enforce the fixed budget via projection.

I. Performance-Weighted Relevance QUBO ("Wrapper" Variant)

To incorporate supervised signals beyond MI, we compute a *performance-weighted* relevance proxy from an inner CV performed on the outer training fold. For each candidate feature i, we estimate its univariate predictive utility via class-weighted logistic regression, producing a score such as inner-fold AUC-ROC. Let $\mathrm{AUC}_i^{(t)}$ denote univariate AUC-ROC on inner split t; we define:

TABLE I. HYPERPARAMETER CONFIGURATION USED IN ALL EXPERIMENTS.

Parameter	Symbol	Value	Scope
Diagonal offset	λ	0.02	All QUBO variants
Redundancy weight	α	0.20	All QUBO variants
Cardinality penalty	μ	1.5	QUBO-k only
Relevance var. penalty	γ	0.25	QUBO-Stability
Redundancy var. penalty	δ	0.10	QUBO-Stability
Wrapper var. penalty	η	0.15	QUBO-Wrapper
Bootstrap replicates	B	40	QUBO-Stability
Inner CV folds	K'	3	QUBO-Wrapper

$$\tilde{r}_i = \frac{1}{K'} \sum_{t=1}^{K'} \mathrm{AUC}_i^{(t)}, \qquad v_i = \mathrm{Var}\big(\mathrm{AUC}_i^{(t)}\big). \qquad (8)$$

We integrate this into the QUBO objective as:

$$E_w(\mathbf{z}) = \sum_{i=1}^{m} \Big(\lambda - \tilde{r}_i + \eta\, v_i\Big) z_i + \alpha \sum_{i<j} \rho_{ij}\, z_i z_j, \qquad (9)$$

with $\eta \geq 0$. We emphasize that this variant uses univariate inner-CV utility as a relevance proxy; it is not a full multivariate wrapper search over subsets. We enforce budget matching via projection.

J. QUBO Optimization via Simulated Annealing

We optimize all QUBO objectives with simulated annealing (SA), a temperature-controlled stochastic search that approximates minima of binary quadratic energy landscapes [9]. The solver returns a binary solution $\mathbf{z}$ and its energy, followed by exact-k projection when applicable.

K. Practical Interpretation of QUBO Hyperparameters

The QUBO coefficients have practical implications for deployment. The parameter λ acts as a diagonal offset, so that when $\lambda < r_i$ the linear term favors selection of feature i, and when $\lambda > r_i$ it penalizes selection. The redundancy weight α controls tolerance to correlated predictors. In QUBO-k, the penalty μ governs cardinality compliance strength. In the stability-regularized variant, γ and δ penalize uncertainty in relevance and redundancy estimates, respectively, improving subset reproducibility at the cost of potentially excluding features whose utility is high but unstable across resamples. In the performance-weighted variant, η trades average utility against robustness. Table I reports all hyperparameter values used.

We fixed all values prior to experimental runs using training-side information only, consistent with the leakage-safe nested CV protocol.

L. Solver Configuration and Reproducibility

Because simulated annealing is stochastic, reproducibility depends on solver configuration choices including temperature schedule, iteration budget, and random seed policy. In this work, we use a standardized SA configuration held constant across all QUBO variants so that comparisons are made under matched procedural conditions. Repeated outer folds partially average stochastic effects, but this should not be interpreted as a full robustness analysis with respect to solver seeds or annealing schedules. A more complete reproducibility study should explicitly vary solver settings and quantify optimizer-induced variance separately from data-resampling variance.

M. Downstream Classifier and Probability Calibration

For fairness across selectors, we evaluate all selected subsets using the same downstream predictor: class-weighted logistic regression producing probabilities calibrated inside the training loop [12], [13]. We report discrimination (AUC-ROC, AUC-PR), classification quality (MCC), and calibration (Brier score). This is important because models with similar AUC can differ materially in probability quality.

N. Evaluation Protocol

We use repeated nested stratified cross-validation to avoid optimistic bias in feature selection and calibration [3], [4]. We run $R = 3$ repeats of $K = 5$ outer folds, yielding $n = 15$ outer test evaluations per method. All selection steps (candidate pooling, bootstraps, and performance-weighted relevance computation) are performed strictly within each outer training fold.

We summarize metrics across the $n = 15$ outer test evaluations using means and 95% confidence intervals. For paired comparisons of AUC-ROC under matched splits, we apply the Wilcoxon signed-rank test [15] (optionally with multiplicity control when many pairwise tests are reported).

Selection stability is quantified by mean pairwise Jaccard similarity across the n selected subsets,

$$J(S_a, S_b) = \frac{|S_a \cap S_b|}{|S_a \cup S_b|}. \tag{10}$$

IV. Results

A. Nested CV Performance

Fig. 1 and Fig. 2 summarize mean AUC-ROC across the repeated nested CV outer tests. Breast Cancer is near-separable for multiple selectors (AUC-ROC close to 1.0), while Pima is moderately challenging (AUC-ROC around 0.83–0.84), where differences are smaller and uncertainty overlaps across strong methods.

B. ROC Curve Analysis (Qualitative)

Fig. 3 and Fig. 4 show ROC curves for qualitative comparison of ranking behavior across thresholds. These plots should be interpreted as illustrative summaries of discrimination; all inferential claims in this study rely on the repeated nested CV outer-test scores reported in Table II and Table III.

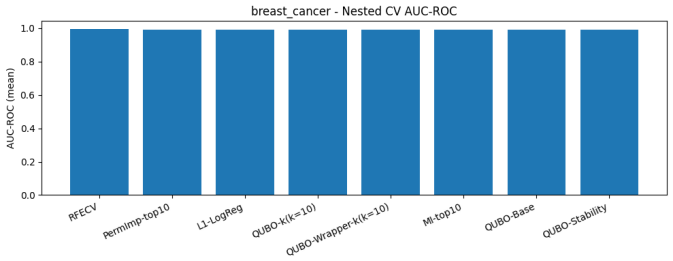

Fig. 1. Breast cancer: Mean nested CV AUC-ROC by method.

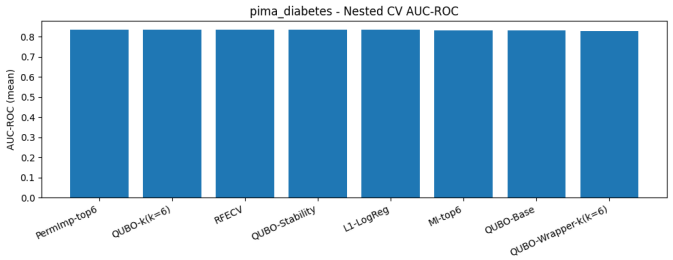

Fig. 2. Pima diabetes: Mean nested CV AUC-ROC by method.

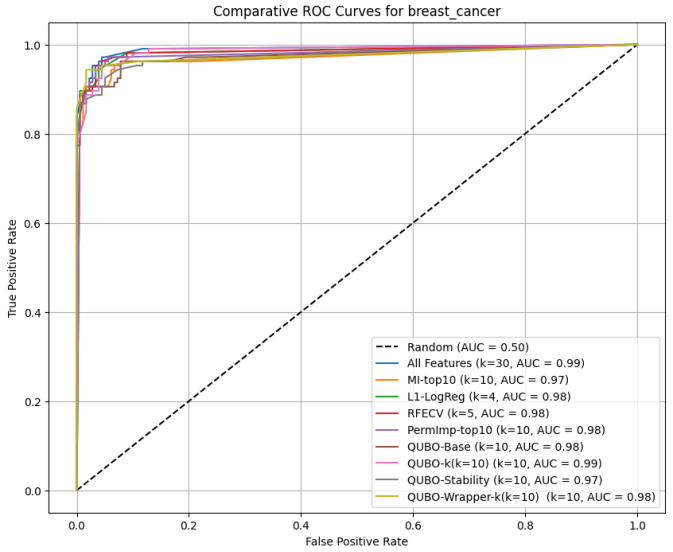

Fig. 3. Comparative ROC curves for breast cancer (illustrative).

For Breast Cancer (Fig. 3), most methods achieve AUC-ROC above 0.98, consistent with a ceiling regime in which marginal AUC differences are difficult to interpret practically. QUBO-k and QUBO-Stability closely track strong baselines (RFECV, L1-logistic), indicating that enforcing combinatorial constraints does not substantially degrade discrimination on this benchmark.

For Pima Diabetes (Fig. 4), separation among curves is more apparent, reflecting a harder decision boundary. QUBO-k remains competitive with permutation-importance and RFECV while enforcing strict cardinality constraints. The performance-weighted relevance variant (labeled "QUBO-Wrapper-k" in tables/figures for consistency with artifact names) shows slightly

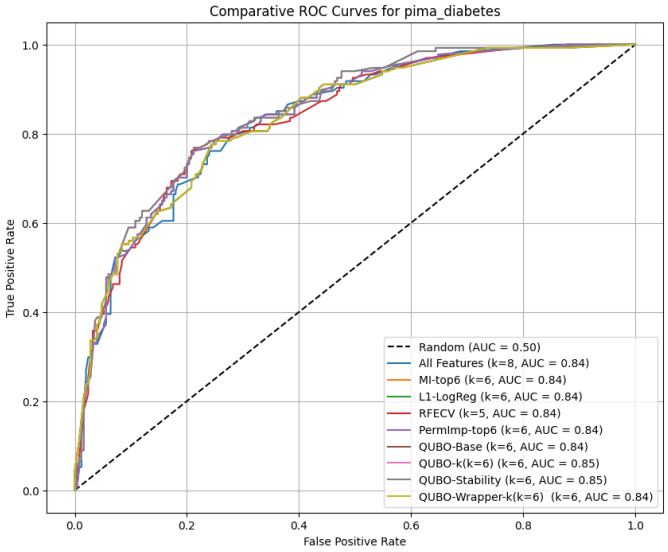

Fig. 4. Comparative ROC curves for Pima diabetes (illustrative).

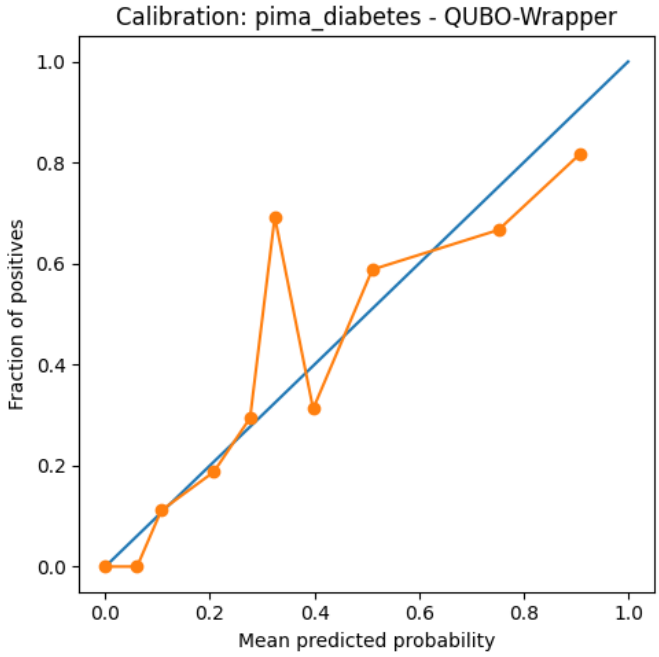

Fig. 5. Calibration curve (reliability diagram) for Pima diabetes using the performance-weighted QUBO variant. The diagonal indicates perfect calibration.

weaker ROC behavior, aligning with the nested CV summaries and suggesting that inner-loop univariate utility signals do not necessarily improve multivariate subset generalization in low-dimensional settings.

C. Calibration

Beyond discrimination, we inspect calibration using Brier score and reliability diagrams. Fig. 5 shows a calibration curve (reliability diagram) for Pima Diabetes for one representative method; deviations from the diagonal indicate over- or under-confidence. Calibration assessment is reported alongside discrimination because near-identical AUC can mask materially different probability quality.

D. Quantitative Summary with 95% Confidence Intervals

Table II and Table III report mean performance and 95% confidence intervals from the repeated nested CV outer tests ($n = 15$). We include discrimination (AUC-ROC, AUC-PR), classification quality (MCC), and calibration (Brier), plus the effective subset size k and selection time. Note that the reported selection times are implementation-dependent wall-clock measurements useful for descriptive comparison under the same environment, but not a fine-grained decomposition of solver versus non-solver cost; values displayed as 0.000 reflect rounding at the reported precision and do not imply zero computational cost.

E. Selection Stability (Jaccard)

Table IV reports mean Jaccard similarity across outer-fold selections. Stability should be interpreted jointly with performance: highly stable selectors are valuable for interpretability, but stability alone can be misleading if performance is low or if the procedure is effectively deterministic for trivial reasons [8].

F. Statistical Comparison

We apply paired Wilcoxon signed-rank tests [15] over the $n = 15$ outer AUC-ROC scores. On Breast Cancer, differences among top-performing methods are not practically distinguishable given overlapping confidence intervals. On Pima, small but consistent differences can appear under matched splits; however, because multiple pairwise comparisons can inflate Type-I error, multiplicity control (e.g., Holm correction) is recommended when reporting many p-values.

V. Discussion

All inferential statements in this section are based exclusively on outer-fold test predictions from the repeated nested CV protocol; ROC and calibration plots are included for qualitative illustration and should not be used as sole evidence of superiority. Nested CV is essential here because feature selection (and any calibration/tuning) is performed on training data, and evaluating on non-isolated splits can inflate performance estimates [3].

A key observation is the presence of a *ceiling regime* on `breast_cancer`. Multiple methods attain AUC-ROC close to 1.0, and confidence intervals overlap. In such cases, it is statistically and practically difficult to justify strong claims based on marginal AUC differences. This does not diminish QUBO; rather, it clarifies its value proposition: QUBO is often most useful when the practitioner must enforce explicit constraints (strict cardinality, domain rules, stability penalties, or cost terms) and wants a transparent optimization objective.

`pima_diabetes` operates in a moderate-difficulty regime where performance is lower and variability across folds is more informative. Here, QUBO-k remains competitive

TABLE II. BREAST CANCER (REPEATED NESTED CV; $n = 15$ OUTER TESTS): MEAN METRICS WITH 95% CI

Method	AUC-ROC [CI]	AUC-PR [CI]	MCC [CI]	Brier [CI]	k (mean)	Sel. time (s)
RFECV	0.9931 [0.9898, 0.9965]	0.9899 [0.9857, 0.9942]	0.9354 [0.9154, 0.9554]	0.0232 [0.0174, 0.0291]	17.93	0.000
PermImp-top10	0.9925 [0.9887, 0.9962]	0.9872 [0.9814, 0.9929]	0.9235 [0.9006, 0.9464]	0.0270 [0.0200, 0.0340]	10.00	0.000
L1-LogReg	0.9923 [0.9880, 0.9966]	0.9893 [0.9838, 0.9947]	0.9331 [0.9153, 0.9510]	0.0244 [0.0179, 0.0309]	9.13	0.000
QUBO-k (k=10)	0.9906 [0.9862, 0.9950]	0.9865 [0.9799, 0.9932]	0.9170 [0.9000, 0.9340]	0.0282 [0.0225, 0.0339]	10.00	1.193
QUBO-Wrapper-k (k=10)	0.9900 [0.9860, 0.9940]	0.9843 [0.9786, 0.9900]	0.9105 [0.8891, 0.9319]	0.0326 [0.0266, 0.0387]	10.00	1.611
MI-top10	0.9900 [0.9864, 0.9935]	0.9852 [0.9806, 0.9897]	0.8981 [0.8733, 0.9230]	0.0375 [0.0308, 0.0443]	10.00	0.000
QUBO-Base	0.9900 [0.9864, 0.9935]	0.9852 [0.9806, 0.9897]	0.8981 [0.8733, 0.9230]	0.0375 [0.0308, 0.0443]	10.00	1.562
QUBO-Stability	0.9893 [0.9833, 0.9954]	0.9862 [0.9790, 0.9933]	0.9308 [0.9081, 0.9536]	0.0273 [0.0193, 0.0352]	10.00	1.648

TABLE III. PIMA DIABETES (REPEATED NESTED CV; $n = 15$ OUTER TESTS): MEAN METRICS WITH 95% CI

Method	AUC-ROC [CI]	AUC-PR [CI]	MCC [CI]	Brier [CI]	k (mean)	Sel. time (s)
PermImp-top6	0.8363 [0.8257, 0.8468]	0.7097 [0.6930, 0.7263]	0.4665 [0.4419, 0.4912]	0.1566 [0.1522, 0.1611]	6.00	0.000
QUBO-k (k=6)	0.8356 [0.8251, 0.8461]	0.7092 [0.6914, 0.7270]	0.4702 [0.4496, 0.4907]	0.1570 [0.1526, 0.1613]	6.00	0.324
RFECV	0.8350 [0.8238, 0.8461]	0.7123 [0.6933, 0.7313]	0.4613 [0.4366, 0.4860]	0.1570 [0.1524, 0.1616]	4.60	0.000
QUBO-Stability	0.8347 [0.8247, 0.8447]	0.7080 [0.6913, 0.7247]	0.4643 [0.4438, 0.4848]	0.1580 [0.1535, 0.1626]	6.00	0.333
L1-LogReg	0.8343 [0.8237, 0.8450]	0.7043 [0.6887, 0.7198]	0.4552 [0.4273, 0.4831]	0.1581 [0.1537, 0.1625]	6.53	0.000
MI-top6	0.8328 [0.8212, 0.8445]	0.7115 [0.6941, 0.7289]	0.4593 [0.4316, 0.4871]	0.1584 [0.1533, 0.1635]	6.00	0.000
QUBO-Base	0.8306 [0.8192, 0.8419]	0.7089 [0.6925, 0.7253]	0.4638 [0.4335, 0.4941]	0.1589 [0.1539, 0.1640]	6.00	0.360
QUBO-Wrapper-k (k=6)	0.8283 [0.8178, 0.8388]	0.7050 [0.6921, 0.7180]	0.4646 [0.4363, 0.4930]	0.1596 [0.1551, 0.1641]	6.00	0.368

TABLE IV. MEAN SELECTION STABILITY (JACCARD) ACROSS OUTER FOLDS

Method	Breast Cancer	Pima Diabetes
QUBO-Base	0.9998	0.7388
QUBO-k	0.5215	0.8721
QUBO-Stability	0.5796	0.7592
QUBO-Wrapper-k	0.6938	0.7748
MI-top$k$	0.9998	0.7456
L1-LogReg	0.8189	0.7912
RFECV	0.5210	0.7894
PermImp-top$k$	0.5952	0.8803

with strong baselines (RFECV and permutation-importance top-k) while guaranteeing exactly $k = 6$ features. This is operationally important when each feature corresponds to acquisition cost or measurement availability.

The performance-weighted relevance variant does not necessarily improve generalization, particularly in low-dimensional settings. A plausible explanation is that univariate relevance can mismatch multivariate subset utility, and inner-loop scoring can inject additional variance that does not translate into better held-out outer-fold performance, a phenomenon consistent with known wrapper sensitivities [10]. This suggests that performance-weighted signals may require stronger regularization or multivariate scoring to consistently outperform simpler baselines.

Stability is crucial when feature selection supports interpretation. Unstable selectors can yield brittle explanations and reduce trust even if AUC is high. On Pima, QUBO-k shows high Jaccard similarity, indicating reproducible subset selection under repeated outer folds. Nevertheless, stability must be interpreted alongside discrimination and calibration [8].

Calibration can diverge from AUC: two models may rank instances similarly (similar AUC) yet produce different probability quality (different Brier/reliability), which matters whenever decisions depend on absolute risk thresholds [14]. Evaluating calibration under the same leakage-safe protocol is therefore essential.

Limitations: The empirical evaluation is restricted to two low-dimensional tabular datasets ($d \leq 30$), one of which exhibits near-ceiling performance. This is sufficient to demonstrate leakage-safe evaluation, budget control, and stability diagnostics, but it does not support strong claims about scalability to high-dimensional regimes (e.g., genomics, text, or finance). Runtime values are descriptive end-to-end measurements and are not decomposed into solver versus non-solver components.

VI. CONCLUSION

This work introduced a stability-aware QUBO-based feature selection pipeline for tabular binary classification and evaluated it under a rigorous repeated nested cross-validation protocol, reporting confidence intervals, calibration-aware metrics, and explicit subset stability. Across the Breast Cancer Wisconsin and Pima Indians Diabetes benchmarks, QUBO formulations achieved discrimination competitive with strong classical baselines (L1-regularized logistic regression, RFECV, mutual-information filtering, and permutation-importance ranking) while additionally providing direct and transparent control over subset cardinality through an explicit k-constraint. Beyond raw accuracy, the results underscore that in near-separable settings where performance approaches a ceiling and differences become statistically and operationally marginal, reproducibility of the selected subsets and the ability to enforce interpretable, budgeted feature sets become the most meaningful differentiators. In this sense, QUBO offers a unified optimization framework to encode relevance–redundancy trade-offs, strict feature budgets, and stability-oriented regularization within a single objective. Future work will extend evaluation to higher-dimensional and cost-sensitive datasets, study solver and compute-budget sensitivity with detailed runtime breakdowns, and explore multi-objective QUBO designs that jointly optimize discrimination, calibration, and stability to further strengthen robustness and interpretability in practical decision-making pipelines.

ACKNOWLEDGMENT

The authors thank the Instituto de Investigación en Inteligencia Computacional y Ciencia de Datos (IIICCD) and the Universidad Nacional del Altiplano for supporting research activities and computational resources.

REFERENCES

[1] A. Lucas, "Ising formulations of many NP problems," *Frontiers in Physics*, vol. 2, p. 5, 2014, doi: 10.3389/fphy.2014.00005.

[2] M. Mücke, N. Piatkowski, and K. Morik, "Feature selection on quantum computers," *Quantum Machine Intelligence*, vol. 5, no. 1, p. 11, 2023, doi: 10.1007/s42484-023-00099-z.

[3] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.

[4] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, no. 1, p. 91, 2006, doi: 10.1186/1471-2105-7-91.

[5] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005, doi: 10.1109/TPAMI.2005.159.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[7] N. Meinshausen and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 417–473, 2010.

[8] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," *Journal of Machine Learning Research*, vol. 18, no. 174, pp. 1–54, 2018.

[9] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983, doi: 10.1126/science.220.4598.671.

[10] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1–2, pp. 273–324, 1997, doi: 10.1016/S0004-3702(97)00043-X.

[11] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002, doi: 10.1023/A:1012487302797.

[12] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schurmans, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.

[13] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, AB, Canada, 2002, pp. 694–699, doi: 10.1145/775047.775151.

[14] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[15] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in Python," arXiv:1201.0490 [cs.LG], 2011.

[17] Scikit-learn Developers, "`sklearn.datasets.load_breast_cancer` documentation," Scikit-learn Stable Documentation, 2025. [Online]. Available: https://scikit-learn.org/

[18] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proceedings of the Symposium on Computer Applications and Medical Care*, Washington, DC, USA, 1988, pp. 261–265.