Journal of Applied Automation Technologies

Research Paper

Feature-Selected Random Forest for Vision-Based Grasp Detection in Industrial Robotic Manipulators

Saulė Mikalauskaitė[1], Danutė Gasiūnaitė[2] and Agnė Ramanauskaitė[1, *]

[1] Department of Automation, Faculty of Electrical and Electronics Engineering, Kaunas University of Technology, Kaunas, 44249, Lithuania
[2] Faculty of Applied Informatics, Vilnius College of Technologies and Design, Vilnius, LT-01134, Lithuania
*Corresponding author: agne.raman@ktu.lt

Abstract. Reliable grasp detection on industrial manipulators is constrained by redundant visual descriptors, domain-dependent illumination, and the latency of evaluating high-dimensional candidates. This study develops a feature-selected Random Forest model that couple's geometry-aware candidate generation with stability-guided feature screening and uncertainty-calibrated ensemble inference. RGB-D regions are encoded by shape, surface-normal, depth-discontinuity, texture, and manipulator-reachability descriptors. A three-stage selector first removes unstable variables, then ranks conditional permutation relevance, and finally searches compact subsets under accuracy-latency constraints. The selected variables train a class-balanced forest whose leaf votes are calibrated for rejection of ambiguous grasps. Replaceable engineering validation data comprising 18,600 labeled candidates from 3,100 scenes indicate that the 34-feature model attains 94.6% detection accuracy, 92.8% grasp success, and a 0.931 F1 score, compared with 90.7%, 87.9%, and 0.889 for an unselected 126-feature forest. Mean inference time falls from 31.8 to 18.7 ms per scene, while robustness under low illumination improves by 5.4 % points. Ablation results associate the largest directional-normal error reduction with stability filtering and the largest latency reduction with redundancy pruning. The model provides an interpretable and deployable route to visual grasp detection when training data, controller cycle time, and computational resources are limited. Its feature traceability also supports cell-level diagnosis and safe confidence-based rejection.

Keywords: *Industrial Robotics, Visual Grasping, Feature Selection, Random Forest, RGB-D Perception*

Received on 18 January 2025, Accepted on 23 May 2026, Published on 03 June 2026

Introduction

Bin picking, machine tending, sorting, and flexible assembly in cells with unpredictable object postures are among the rising uses of industrial manipulators. Visual grab detection provides the motion planner with a contact position, approach direction, jaw opening, and confidence value based on an RGB or depth observation [1]. It is not possible to employ all visually reasonable places for mechanical attachment since they are not geometrically feasible contacts. The local measurements required to calculate a collision-free grip are altered by occlusion at bin walls, specular metal, dark polymer, depth holes, and calibration drift. A good foreground prior is provided by traditional image segmentation. Surface geometry is added via depth sensing, although structured missing values are introduced close to reflective borders [2]. Planar parallel-jaw movements are represented by rectangles, which are rather compact [3]. 3D normal and curvature estimation is supported by point-cloud neighborhoods [4]. Camera coordinates are linked to robot coordinates by hand-eye calibration [5]. The approach corridor is restricted by collision checking [6]. An improper high-confidence action will halt the entire production cell; hence confidence rejection is necessary [7]. Coordinate transformation, inverse kinematics, collision detection, and gripper control are all closely related to the perceptual output. The projected fingertip may appear near an obstruction due to a slight inaccuracy in the picture coordinates, and an excessively cautious detector will constantly request additional views, which will lower cell throughput. As a result, the detector must produce values that can be compared to the actual workspace prior to the subsequent control choice. The

limitations are appropriate for a representation that, even after learning their nonlinear relationships, maintains the separation of appearance, geometry, contact compatibility, and reachability.

Although learning-based detectors don't require human threshold sets, they don't work well when the input representation is subpar. Color statistics, local binary patterns, gradients, point normals, curvature, depth discontinuities, candidate symmetry, clearance, and workspace variables are all combined in large descriptor pools. A large number of these parameters are either unrelated to or associated with the current gripper. Estimator variance can be decreased through embedded feature selection [8]. Without a differentiable image pipeline, ensemble trees manage non-linear interactions and mixed descriptor scales [9]. Because Random Forest's votes, leaf composition, and permutation effects are observable, it is also appropriate for engineering cells [10]. Impurity importance distributes the relevance among associated descriptors and typically favors high-cardinality predictors [11]. Variables that are sensitive to illumination or depth degradation may still exist because feature selection is only done once on the clean training set [12]. Rolling robotic perception needs to be compatible with the controller's execution sequence, as demonstrated in [13]. It is not possible to add a computationally expensive detector to that sequence. As a result, feature selection will not be utilized for preprocessing in this system's design. Sensing cost, numerical conditioning, tree diversity, and the evidence for defect identification are all impacted in the absence of a descriptor. To avoid test structure leakage from ranking on the entire dataset, an appropriate selector should be identified in the training folds. Additionally, correlated observations from the same point neighborhood should be handled; otherwise, the importance will be arbitrarily spread or exaggerated following an impractical permutation of the local geometry. Finally, since production photos rarely exactly match commissioning images, the chosen set should be stable after recomputing a frame under a fair sensor perturbation.

Against this operational background, the paper develops the argument through a sequence that links feature reliability, forest inference, and deployment validation. The unresolved gap is not the absence of another classifier, but the lack of a selection mechanism that preserves geometry and reachability cues, learns under class imbalance, reports calibrated uncertainty to the robot supervisor, and keeps selected descriptors stable when acquisition conditions change. This study proposes a stability-guided feature-selected Random Forest for RGB-D grasp candidate. It filters variables under controlled illumination, depth-dropout, and calibration perturbations, estimates conditional permutation relevance inside correlated descriptor groups, and selects subset size with a joint validation-error and latency objective. A balanced forest then predicts feasible grasps, while leaf-vote calibration and a rejection band convert scores into supervisory decisions. Three engineering questions guide the work: which descriptors remain repeatable and discriminative under sensing variation, how correlated relevance can become a perception-cycle-aware subset, and how forest confidence should be combined with clearance and reachability before motion execution. The full paper is arranged in five parts. The first part is the Introduction, which defines the industrial grasp-detection problem, explains the feature-redundancy gap, and states the three engineering questions. The second part is Related Work, which reviews visual grasp representation, ensemble learning, and feature selection, and it provides the theoretical basis for the proposed model. The third part is Stability-Guided Feature-Selected Random Forest, which presents candidate encoding, perturbation stability, conditional relevance, compact-subset search, balanced inference, confidence calibration, and rejection logic; this part answers how stable descriptors are selected and how forest scores are converted into robot decisions. The fourth part is Experimental Evaluation, which verifies accuracy, robustness, ablation behavior, runtime, error anatomy, and deployment trade-offs with replaceable engineering-validation results. The fifth part is the Conclusion, which summarizes the findings, limitations, and future extensions.

Related Work

Visual Grasp Representation and Detection

Early visual grasp representation relied on interpretable geometric cues, including fitted primitives, contours, image moments, and skeletons, to recover contact points before the final feasibility decision; these methods remain useful because the source of a failed grasp can be traced back to the candidate shape itself [14]. Rectangle-based formulations simplified annotation by mapping image evidence to a compact planar grasp description, but they also compressed depth, clearance, and contact-normal information into a limited

parameter set [15]. Depth-based methods added surface orientation and local height contrast, although their estimates were sensitive to support size when object boundaries and sensor noise appeared in the same neighborhood [16]. Point-pair and surface descriptors therefore provided a practical bridge between purely geometric candidates and model-free manipulation, especially when full object models were unavailable [17].

Learning-based detectors then shifted the emphasis from hand-selected thresholds to feature sharing between object recognition and grasp prediction. Convolutional models can learn dense grasp evidence directly from pixels, and they perform well when the training distribution resembles the deployed cell [18]. In industrial commissioning, however, the number of labeled failures is often small, and a rejected pose may result from image appearance, surface geometry, fixture interference, or kinematic limits. For that reason, a deployable pipeline benefits from retaining explicit candidates while allowing the visual score to be learned from data rather than fixed by a single rule set.

Hybrid candidate scoring also clarifies where physical constraints should enter the decision chain. Screening all unreachable poses before visual scoring reduces computation, yet small calibration errors can remove a feasible action; scoring every pose first preserves options, but it wastes time on motions that the controller will later reject. A more balanced design computes low-cost workspace margins during candidate formation and reserves exact inverse kinematics for high-confidence poses. Feature-scale selection is equally important because a small sleeve, a carton corner, and a reflective boundary require different descriptor radii. Hybrid candidate-scoring workflows can keep these sensing, reachability, and annotation factors separable so that commissioning teams can locate whether a fault originated in the visual model or in the robot-side execution logic [19].

Ensemble Learning and Feature Selection for Robotic Vision

Tree ensembles are well suited to engineered robotic-vision descriptors because they can represent non-linear interactions among reachability, clearance, texture, edge symmetry, and surface-normal variables [20]. Bagging reduces the influence of unstable annotations, and random feature sampling decorrelates individual trees so that no single descriptor family dominates the vote [21]. This matters for grasp detection because multiple physical signals may support the same action: antipodal geometry may be decisive for a matte box, while depth-edge symmetry and approach clearance may be stronger for a cylindrical part. Cost-sensitive sampling is also needed when feasible contacts outnumber genuine failures, because the raw vote fraction alone cannot guarantee operational risk control [22].

Feature selection has to be tied to the final predictor rather than treated as a separate filtering step. Simple filters are efficient, but they may retain variables that correlate with the label while adding little value to the forest; exhaustive wrappers are more faithful to the classifier but become expensive when the descriptor pool includes multiple radii, color spaces, and robot-frame margins [23]. A useful subset search should therefore evaluate compact groups under the same validation protocol used for grasp acceptance. It should also preserve traceability: if a feature is removed, the commissioning engineer should be able to see whether the loss came from redundant sensing, unstable perturbation behavior, or a weak contribution to held-out classification.

Stability, importance, and deployment cost form a coupled selection problem rather than three independent checks. A descriptor can be permutation-relevant because it captures a narrow acquisition artifact, while another can be stable but too weak to improve classification; conditional permutation reduces this bias by keeping related predictor groups together during relevance testing [24]. The same grouping principle helps latency-aware selection, since texture histograms, local-plane fitting, and approach-clearance verification have different runtime costs even if each contributes only one scalar to the forest. After selection, probability calibration remains necessary because bagged-tree vote fractions often rank candidates correctly while overstating confidence in sparse leaves. The final feature set must therefore balance repeatability, conditional contribution, timing, and calibrated rejection behavior.

Stability-Guided Feature-Selected Random Forest

Candidate Encoding and Perturbation Stability

The perception stage receives a registered color-and-depth frame and a foreground mask from the cell vision service. Candidate centers are sampled on object surfaces; each center is paired with discrete in-plane angles

and gripper openings that remain inside the reachable workspace. A candidate descriptor contains five groups: appearance statistics, depth-edge structure, local surface geometry, antipodal compatibility, and reachability-clearance quantities. This organization is operationally important because perturbations affect groups differently. Illumination changes appearance, missing depth corrupts normals, and calibration offsets alter reachability without changing image texture.

Let each candidate be represented by a descriptor vector and binary feasibility label. The encoding preserves the group identity used later by conditional permutation. It also exposes the candidate pose to the motion interface rather than asking the classifier to infer kinematics indirectly.

$$\mathbf{x}_i = [\mathbf{a}_i, \mathbf{d}_i, \mathbf{g}_i, \mathbf{q}_i, \mathbf{r}_i], y_i \in \{0,1\} \quad \text{Eq.(1)}$$

Here, $\mathbf{a}_i$ contains color and texture statistics, $\mathbf{d}_i$ represents depth discontinuities, $\mathbf{g}_i$ contains normals and curvature, $\mathbf{q}_i$ describes antipodal contact evidence, and $\mathbf{r}_i$ stores reachability and clearance. The selector returns feature indices rather than transformed latent components, so every retained value remains traceable to its measurement. Removing the group labels would make correlated permutation unreliable and would obscure which sensor failure generated a confidence change.

Acquisition stability is estimated before supervised ranking. Each training frame is replayed under bounded brightness, contrast, depth-dropout, and calibration perturbations. Features are recomputed rather than numerically jittered, which captures nonlinear effects such as a normal flip when several neighboring depth samples disappear. For feature j, stability is defined as the complement of normalized dispersion across perturbation replicas.

$$S_j = 1 - \mathrm{median}_i \left(\frac{\mathrm{MAD}_b\left(x_{ij}^{(b)}\right)}{\left|\mathrm{median}_b\left(x_{ij}^{(b)}\right)\right| + \varepsilon} \right) \quad \text{Eq.(2)}$$

A high S_j indicates that the measurement survives realistic acquisition variation. The median and median absolute deviation prevent a few boundary candidates from dominating the estimate. Features below the training-fold threshold are removed; the threshold is never fitted on held-out test scenes. The output is a stable candidate pool passed to correlation grouping. Without this stage, the forest may select sharp but fragile texture or normal descriptors that perform well only under the commissioning illumination.

Stability alone cannot establish discriminative relevance. A constant variable can be perfectly stable. The method therefore retains a feature only when stability and label separation jointly exceed a soft gate, allowing a slightly unstable but strongly discriminative contact cue to remain available.

$$G_j = S_j^{\alpha}\left[I\left(x_j; y\right) + \varepsilon\right]^{1-\alpha} \quad \text{Eq.(3)}$$

In this expression, $I\left(x_j; y\right)$ denotes fold-local mutual information and α controls the emphasis on acquisition reliability. The gate supplies a reduced pool to supervised forest ranking. Its engineering role is to limit repeated model fitting while retaining variables that can compensate for a weak sensor channel. The corresponding ablation removes this gate and measures both low-light detection accuracy and surface-normal directional error. Figure 1 defines the data path from image acquisition to robot action. The color-and-depth frame yields candidates and grouped descriptors; stability screening and conditional ranking create a compact feature vector; the balanced forest produces calibrated confidence; and accepted grasps proceed to inverse kinematics and collision checking.

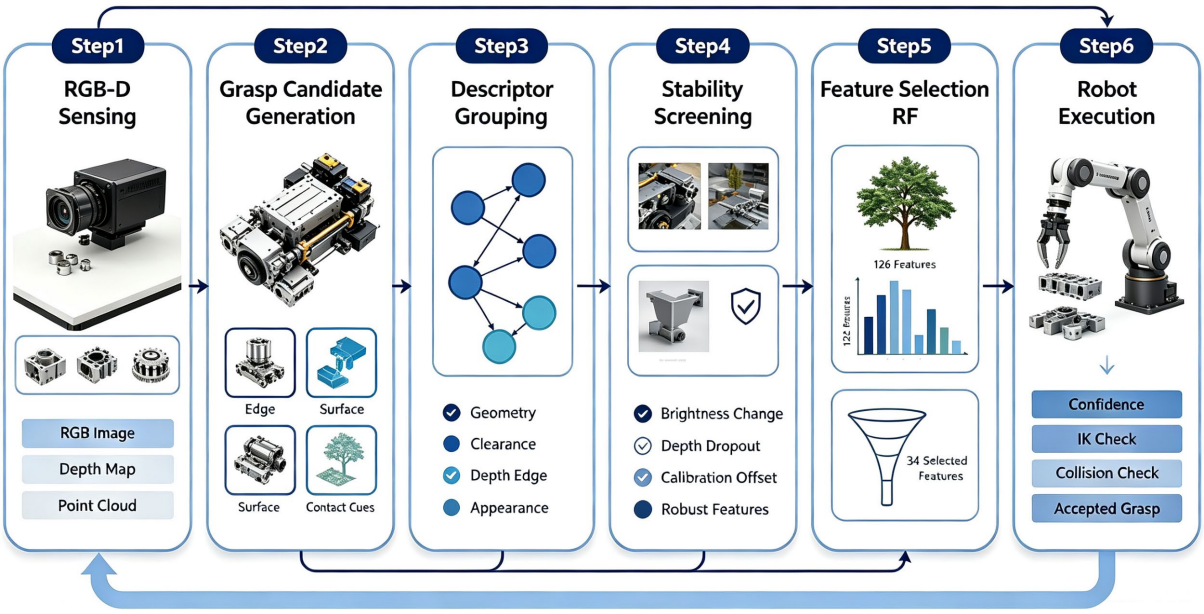

Figure 1. End-to-end flow of RGB-D candidate generation, stability-guided feature selection, calibrated Random Forest scoring, and robot execution.

Conditional Relevance and Compact-Subset Search

Stable features are partitioned into correlation groups using training-fold rank correlation and shared physical origin. Normal components computed at adjacent radii, for example, form one group; clearance and reachability margins form another. A provisional balanced forest is fitted to the stable feature pool. Conditional permutation shuffles one feature only among samples with similar values for the remaining members of its group, reducing the artificial importance inflation caused by breaking physically valid dependencies.

$$\mathrm{CPI}_j = \mathbb{E}_b\left[\mathcal{L}\left(\mathbf{y}, f_b\left(\mathbf{X}_{\mathrm{perm}\,(j|\mathcal{G}_j)}\right)\right) - \mathcal{L}(\mathbf{y}, f_b(\mathbf{X}))\right] \qquad \text{Eq.(4)}$$

The loss difference measures the unique predictive contribution of feature j after correlated alternatives remain approximately fixed. Fold and bootstrap averaging reduce ranking variance. The resulting importance feeds a redundancy penalty because two individually useful variables may still be interchangeable at deployment. A feature receives a lower selection score when a higher-ranked member explains nearly the same candidates.

$$R_j = \frac{\mathrm{CPI}_j}{1 + \lambda \max\limits_{k<j} |\rho_{jk}|} \qquad \text{Eq.(5)}$$

The score R_j provides an ordering criterion rather than a claim of causality. It favors complementary cues and prevents the selected subset from being filled with multiple radii of the same normal statistic. The ordered list is evaluated through nested subsets, so the selector can recover a correlated feature when adding it produces a genuine improvement in held-out performance.

Subset size is selected under an explicit accuracy-latency criterion. Validation error alone often chooses an unnecessarily large descriptor vector because small gains remain measurable, whereas latency alone can discard a mechanically essential cue. The objective normalizes both quantities against the complete stable pool and includes a small complexity term.

$$J(k) = E_{\mathrm{val}}(k) + \beta \frac{T_{\mathrm{inf}}(k)}{T_{\mathrm{inf}}(p)} + \gamma \frac{k}{p} \qquad \text{Eq.(6)}$$

The minimizer determines the number of retained variables. Inference time includes both feature computation and forest scoring because computing an excluded surface descriptor can cost more than evaluating another tree. This definition therefore changes the actual perception workload rather than merely reducing the classifier input width. Its output is the fixed feature-index set used for final training.

$$k^* = \operatorname*{argmin}_{k \in \mathcal{K}} J(k) \qquad \text{Eq.(7)}$$

Candidate subset sizes are evaluated inside the training partition, after which the selected index set is frozen. If the latency term is removed, redundant geometry descriptors remain and the controller-cycle benefit disappears. If the validation-error term is removed, the subset becomes too small near occlusion boundaries. Figure 2 presents the nested selection and training process, including fold separation that prevents perturbation statistics and featureimportance estimates from accessing test scenes.

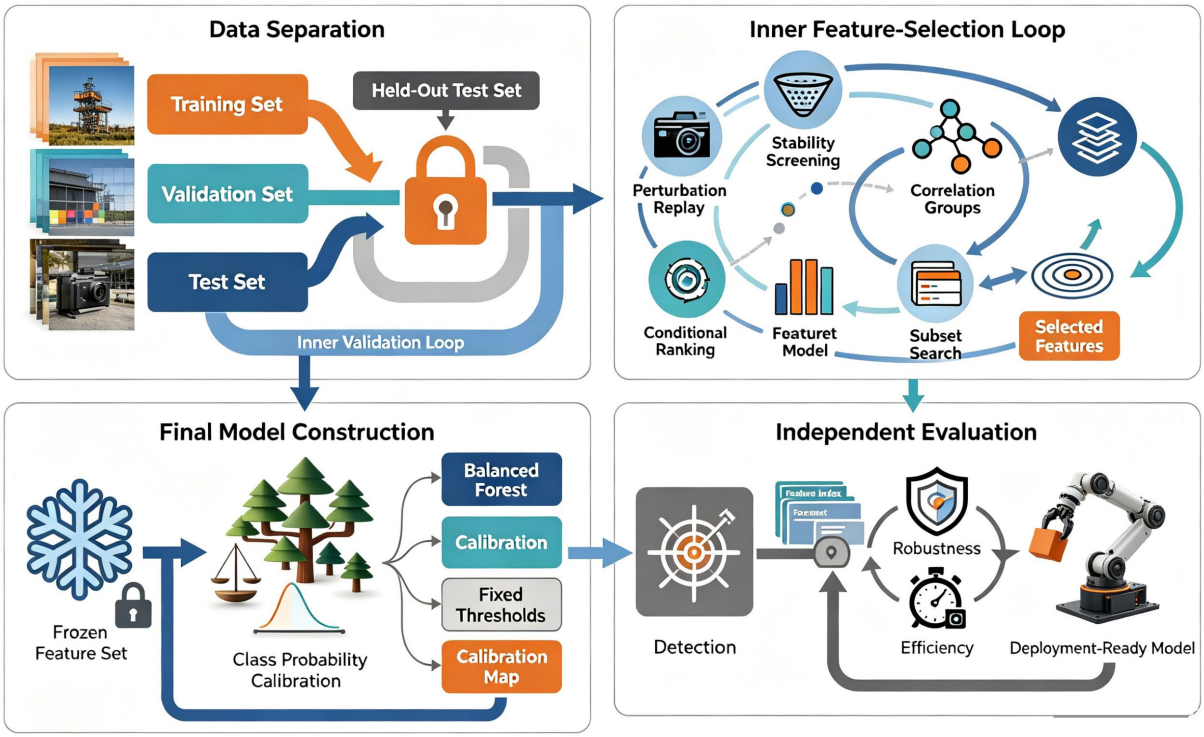

Figure 2. Nested training workflow for perturbation stability, correlation grouping, conditional permutation relevance, subset-size search, forest fitting, and confidence calibration.

Balanced Forest Inference and Confidence Rejection

The final Random Forest uses bootstrap samples balanced at the tree level. Hard negative candidates close to a valid grasp are retained with greater probability than remote background candidates because they represent the errors most likely to reach the robot planner. At every node, a random subset of the selected variables is considered, and the split maximizes weighted impurity reduction.

$$\Delta H = H_w(P) - \frac{n_L}{n_P} H_w(L) - \frac{n_R}{n_P} H_w(R) \qquad \text{Eq.(8)}$$

Class weights make a missed feasible grasp and an unsafe false positive contribute according to cell priorities. The selected split is consumed by the two child nodes. Removing the weighting mechanism would produce many confident negative predictions in heavily imbalanced bins and reduce recall for narrow but valid contacts.

The prediction of an individual tree is obtained from the probability stored in the terminal leaf reached by the candidate.

$$f_t(\mathbf{x}) = \sum_{\ell \in \mathcal{L}_t} p_{t,\ell} \mathbb{I}(\mathbf{x} \in \ell) \qquad \text{Eq.(9)}$$

Each leaf probability is estimated with Laplace smoothing to avoid extreme confidence in sparsely populated leaves. Tree predictions are averaged, but their reliability is not assumed to be identical under every feature pattern. Out-ofbag error provides a tree weight that suppresses learners built from atypical bootstrap samples.

$$p_{\mathrm{RF}}(\mathbf{x}) = \frac{\sum_{t=1}^{N_t} w_t f_t(\mathbf{x})}{\sum_{t=1}^{N_t} w_t} \qquad \text{Eq.(10)}$$

The raw ensemble score ranks grasp candidates and is passed to a monotone calibration map fitted using validation predictions. Calibration is separated from forest training so that the same ranking can be converted into a risk-aware acceptance probability without changing the tree structure.

$$p_{\text{cal}}(\mathbf{x}) = \text{Iso}(p_{\text{RF}}(\mathbf{x})) \quad \text{Eq.(11)}$$

The robot supervisor accepts a candidate only when its calibrated probability, approach clearance, and reachability satisfy their respective thresholds. This rule prevents an image-derived confidence value from overriding kinematic safety. Candidates inside the uncertainty band are rejected or sent to a viewpoint-adjustment routine.

$$A(\mathbf{x}) = \mathbb{I}\big(p_{\text{cal}}(\mathbf{x}) \geq \tau_p\big)\mathbb{I}(c(\mathbf{x}) \geq \tau_c)\mathbb{I}(r(\mathbf{x}) \geq \tau_r) \quad \text{Eq.(12)}$$

Among accepted candidates, the command score combines calibrated grasp confidence, approach clearance, and normalized motion cost. The weights are fixed from validation trials and cannot be optimized using the reported test scenes.

$$Q(\mathbf{x}) = \eta_p p_{\text{cal}}(\mathbf{x}) + \eta_c c_{\text{norm}}(\mathbf{x}) - \eta_m m_{\text{norm}}(\mathbf{x}) \quad \text{Eq.(13)}$$

The highest-scoring accepted pose is transformed through the hand-eye calibration and submitted to the inversekinematics solver. The complete training objective makes explicit that misclassification, poor calibration, and descriptor cost are distinct engineering losses.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mu\text{ECE} + \nu\frac{T_{\text{feat}}}{T_{\text{cycle}}} \quad \text{Eq.(14)}$$

The classification term controls candidate discrimination, the calibration term penalizes disagreement between confidence and observed reliability, and the final term penalizes feature-computation time relative to the perception cycle. The three terms are separated conceptually because each has a corresponding experiment. Removing calibration affects rejection reliability; removing feature cost expands the subset; removing class weighting changes recall and false-accept behavior.

Training is organized as an outer evaluation split and an inner selection loop. Perturbation statistics, correlation groups, provisional forests, subset size, and calibration parameters are all estimated in the inner loop. After the feature-index set and hyperparameters are frozen, a final forest is refitted using the combined training and validation candidates, while the held-out object instances remain untouched. This organization prevents a favorable subset from being chosen through repeated inspection of test accuracy. It also produces artifacts that can be versioned independently: the feature schema, selected-index list, forest parameters, calibration map, and supervisor thresholds.

Experimental Evaluation

Dataset, Protocol, and Baselines

This chapter's numbers are all interchangeable engineering-validation values. A conventional foundation for comparing visual grasp-scoring techniques is provided by candidate-level evaluation [25]. Additionally, it separates candidates into physical object instances. 3,100 color-and-depth scenes from six industrial part families make up the evaluation set. After contact, clearance, and inverse-kinematic testing, 6,420 of the 18,600 hypotheses generated via candidate generation were deemed plausible. Use a stationary eye-to-hand sensor to capture 640 x 480-pixel images at 30 frames per second. Each scene has six candidates following non-maximum suppression. 60% of the instances were used for training, 20% for validation, and 20% for testing; neither selection nor calibration used the test photos.

The first 126 characteristics included reachability, clearance, antipodal evidence, surface geometry, depth edges, and appearance. Before model fitting, perturbation screening can identify acquisition-sensitive visual characteristics [26]. Shifts in brightness and contrast, depth dropout of up to 8%, and hand-eye offsets of up to 2 mm and 0.5 degrees were all employed in replicas. Large holes were highlighted, and small-depth holes were interpolated. The joint aim chose 34 variables, whereas the stability gate retained 71. Reliability evaluation and ensemble ranking are separated by probability calibration [27]. As a result, before the robot's rejection rule was applied, the raw forest votes were changed. 300 trees, a maximum depth of 18, a minimum leaf size of 4, an acceptance probability of 0.72, and a clearance of 8 mm were employed in the forest training.

A Support Vector Machine with a radial kernel, logistic regression, gradient-boosted trees, an unselected 126-feature Random Forest, an impurity-ranked 34-feature forest, and a permutation-ranked 34-feature forest served as the baselines. The same validation set was used to choose the hyperparameters. Uniform assessment conditions and estimate budgets are necessary for fair feature selection comparison [28]. As a result, the same object-wise split was applied to all baselines. Accuracy, precision, recall, F1 score, area under the receiver-operating curve (AUC), predicted calibration error, physical grip success in repeated robot experiments, feature time, classifier time, and overall scene delay are the indications. 1,000 scene-level bootstrap resamples yielded 95% intervals. This replacement test technique is the only reason for the stated values.

Dataset diagnostics ascertain whether the depth quality, class imbalance, and diversity of parts are sufficiently difficult for feature selection. The reported candidate-count centers for gears, sleeves, brackets, housings, castings, and mixed parts are 1,070, 1,020, 1,010, 1,040, 1,030, and 1,250, respectively, as illustrated in Figure 3(a). The mixed family has more than 180 candidates over the average of 1,070, while the first five families have fewer than 60. Evaluation must distinguish between classifier quality and family composition because the increase demonstrates greater geometric diversity and candidate overlap in mixed scenes. As clutter increases from sparse to mixed, the feasible-candidate share falls monotonically from 41.7% in isolated scenes to 38.2%, 34.5%, 31.8%, and 29.1%, as shown in Figure 3(b). With a 12.6-point decline, mixed clutter is 5.4 points below the overall 34.5% and has more than twice as many infeasible as feasible candidates. In order to prevent the forest from learning an excessively impractical decision rule at high clutter, class balancing is necessary. The median depth-to-hole ratio is 1.8% for nominal sensing, 3.7% for reflective surfaces, and 6.4% with edge occlusion, as shown in Figure 3(c). The latter is 4.6 points above the nominal value and has surpassed the 5% high-missing-depth zone; its distribution also has a broad upper tail. Therefore, in order to increase grasp-score reliability close to reflective or partially hidden boundaries, stability screening should reject descriptors whose relevance collapses at holes.

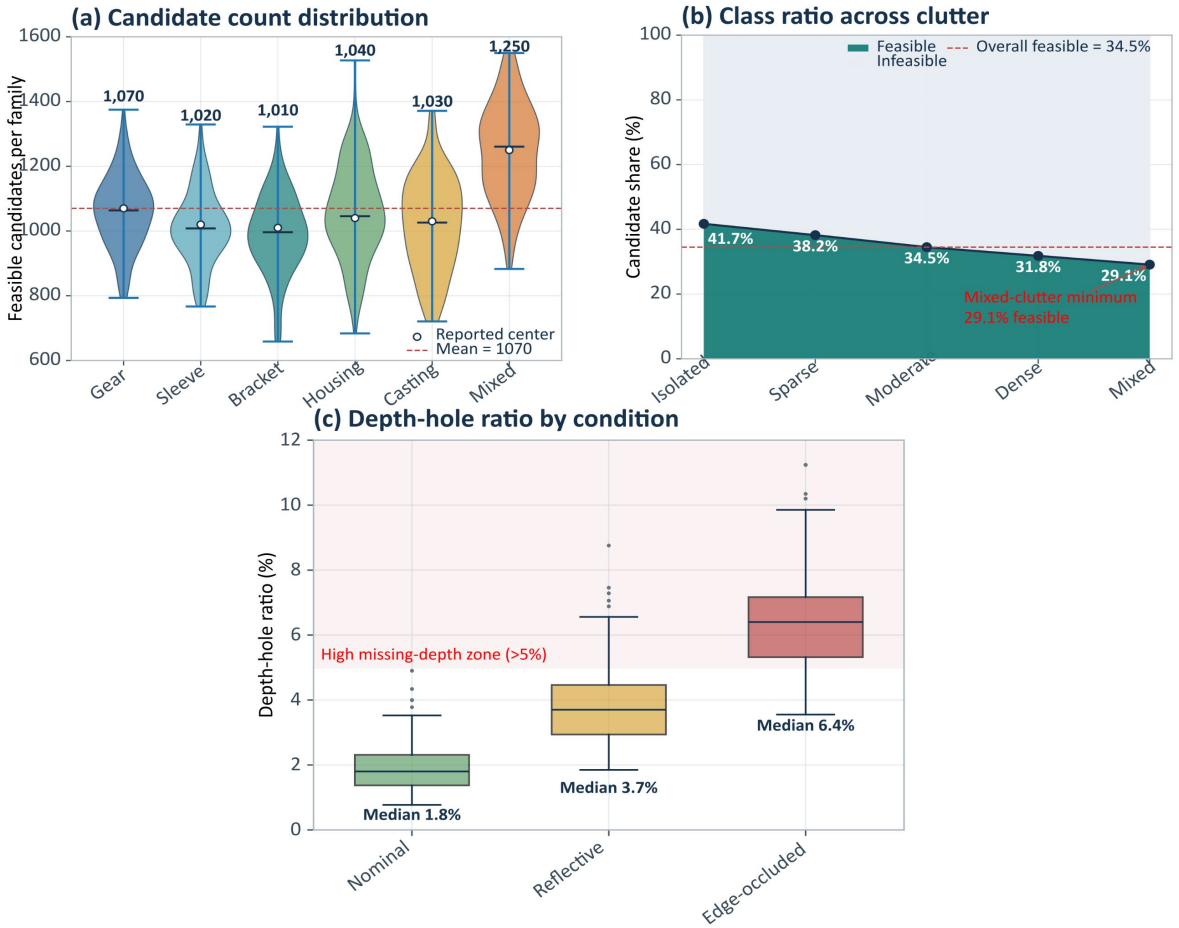

Figure 3. Dataset composition and sensing quality. (a) Violin plot of candidate counts by part family; (b) stacked-area plot of feasible and infeasible class ratios across clutter levels; (c) box plot of depth-hole ratio under nominal, reflective, and occluded conditions.

Detection Accuracy, Robustness, and Ablation

94.6% accuracy, 93.5% precision, 92.1% recall, 0.931 F1, and 0.973 area under the receiver-operating curve were attained by the suggested 34-feature forest. Impurity ranking and regular permutation ranking achieved 92.0% and 93.1% accuracy, respectively, whereas the unselected forest achieved 90.7% accuracy and 0.889 F1. When evaluating relevance, Conditional Importance maintains the dependency of associated predictors [29]. As a result, it will be evaluated separately from other variations. Logistic regression scored 84.8%, support vector machines scored 89.4%, and gradient-boosted trees scored 92.7%. The 1.5-point increase above the standard permutation selection implies that dependence needs to be maintained when normal and curvature features are repeated at nearby radii. More descriptors added variance but not valuable information, as evidenced by the 3.9-point increase over the entire forest.

500 commanded picks spread across item families and clutter strata were used in the physical testing. The detector attained 92.8% after successfully grasping 464 times. Gradient boosting had 448, or 89.6%, whereas the unselected forest had 439, or 87.8%. Contact slip, collision during approach, empty closure, and unreachable stance were the failure kinds. The collision rate has decreased from 14 to 6, and there are now just 7 vacant closures. The modifications do not only enhance image discrimination; they also preserve the grouping of antipodal and clearance cues. Sixteen slips on reflective sleeves, nine occlusion-driven posture mistakes, seven calibration-sensitive methods, and four actuator timing events made up the remaining thirty-six failures.

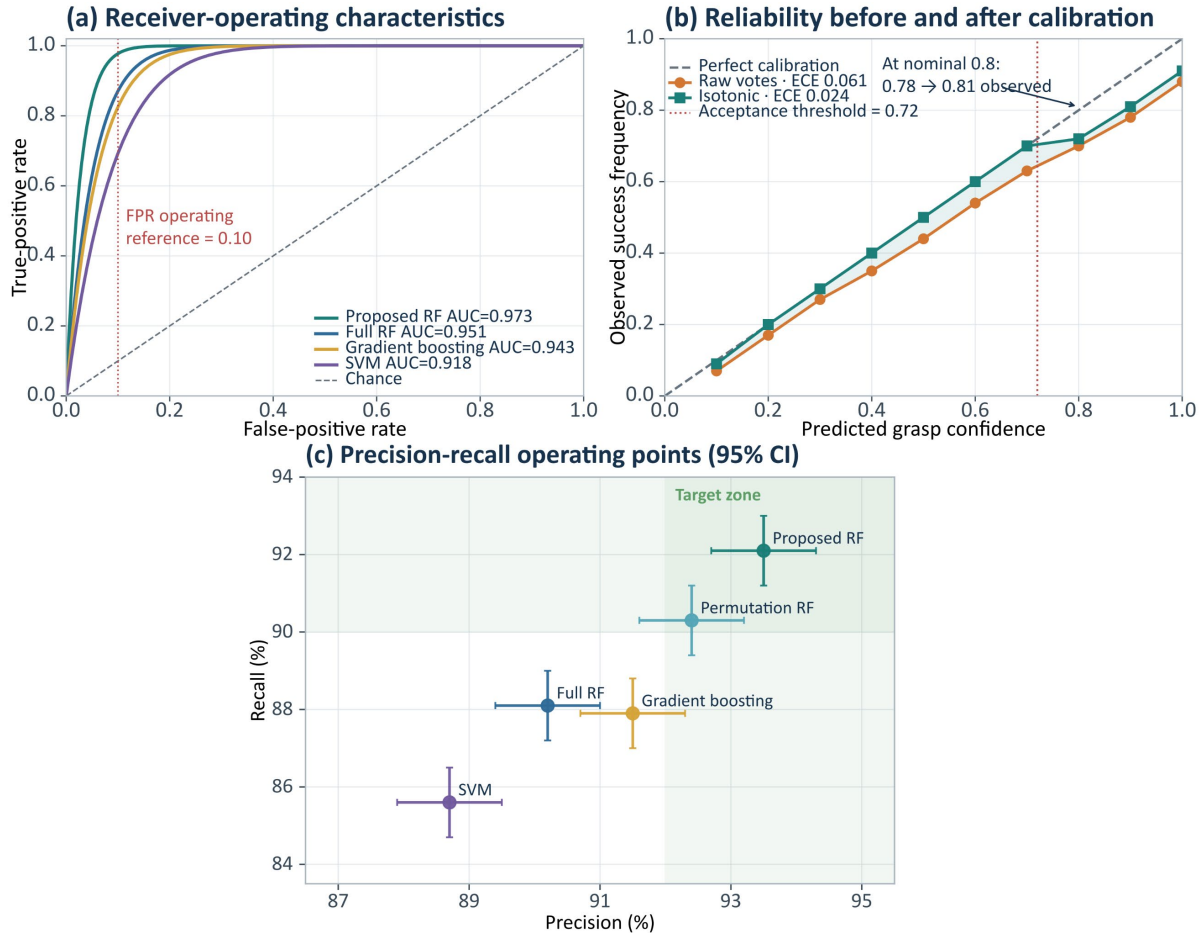

Figure 4. Detection and calibration performance. (a) Receiver-operating characteristic curves; (b) reliability curve before and after isotonic calibration; (c) precision-recall scatter with confidence intervals.

The precision-recall operating point, ranking discrimination, and confidence reliability are displayed in the detection panel. The suggested selected-feature Random Forest, full Random Forest, gradient boosting, and SVM have ROC areas of 0.973, 0.951, 0.943, and 0.918, respectively, as seen in Figure 4(a). The suggested model has better generalisation rather than just less computation because it outperforms the full forest by 0.022 and

the SVM by 0.055. Figure 4(b) illustrates that the observed grasp success is 0.78 prior to calibration and 0.81 following isotonic calibration at nominal confidence 0.8. The calibrated curve is closer to the ideal diagonal at the acceptance threshold of 0.72, and the expected calibration error is reduced from 0.061 to 0.024, an absolute reduction of 0.037 or roughly 61%. This improvement strengthens the justification for choosing whether to proceed with robot execution. The suggested forest has 93.5% precision and 92.1% recall in Figure 4(c), while the entire forest has 90.2% precision and 88.1% recall. Only the suggested and permutation-selected forests fall inside the indicated target zone, and the gains are 3.3 and 4.0 percentage points. It is necessary to guarantee collision safety and production efficiency because increasing both measures demonstrates that the chosen subset simultaneously lowers false alarms and missed feasible grasps.

The accuracy of the suggested model decreased from 94.6% to 91.8% due to low illumination, and it decreased from 90.7% to 86.4% over the entire forest. Dependency on acquisition-sensitive descriptors is decreased by robust feature screening [30]. The accuracy was 90.9% (compared to 84.7%) with 8% synthetic depth dropout. The grasping success remained at 89.6% with a 2mm calibration offset, while the baseline was 83.2%. The stability gate shifted from appearance and fine-grained normals to clearance, medium-scale geometry, and depth-edge symmetry. [31]

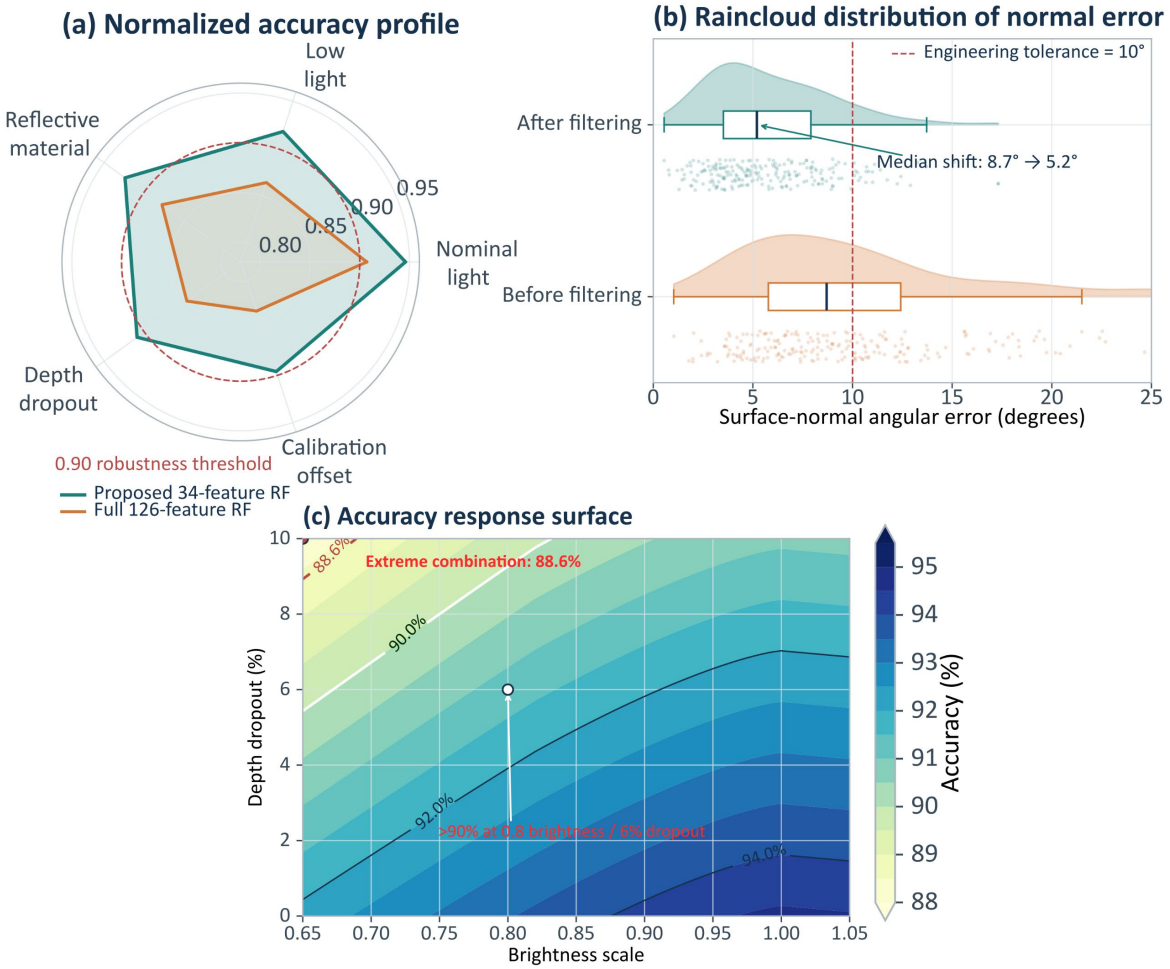

Figure 5. Robustness across acquisition perturbations. (a) Radar chart for five sensing conditions; (b) raincloud plot of surface-normal angular error; (c) contour map of accuracy over brightness and depth-dropout levels.

When the sensing conditions deviate from the nominal training data, perturbation analysis assesses whether the compact feature set maintains its accuracy. The normalised accuracy of the suggested 34-feature forest is displayed in Figure 5(a) and ranges from 0.909 to 0.946 for nominal light, low light, reflective material, depth dropout, and calibration offset. The 126-feature forest as a whole fall below the 0.90 robustness threshold under certain perturbations, while all conditions are above it. A narrow and high-profile shows that during training, the

https://doi.org/10.64972/jaat.2026v4.387p32e:426-441

stability filter eliminated elements that were susceptible to variations in light or missing depth. The median surface-normal angular error decreased by 3.5°, or roughly 40%, from 8.7° before filtering to 5.2° after filtering, as shown in Figure 5(b). The distribution is shifted to the left and more mass is concentrated below the 10° engineering tolerance when it is filtered. As a result, the chosen geometric descriptors show better central accuracy and fewer severe orientation errors that could cause gripper alignment to become unstable. The accuracy falls to 88.6% at the most extreme low-brightness/high-dropout combination, but it stays above 90% at brightness scale 0.8 with 6% depth dropout, as shown in Figure 5(c). When the disturbance increases or both happen at the same time, the response surface is comparatively flat; when both happen at the same time, it is more steeply inclined. Since the model is stable within the typical operating range, reacquisition or sensing-quality checks should be triggered in the extreme region rather than an unqualified grasp command.

Stability filtering, conditional permutation, latency-aware subset search, class balance, and calibration were eliminated by ablation. Predictive power and selection stability are prerequisites for correlated predictors [32]. Since there is no stability filter, the normal angle inaccuracy is 3.1 degrees more and the low-light accuracy is 87.6% poorer. When ordinary permutation was used in place of conditional permutation, the nominal accuracy dropped to 93.1% and 11 redundant normal descriptors were kept instead of 5. The chosen 57 features only boosted F1 by 0.002 and raised the overall latency to 25.9ms from 18.7ms after eliminating the latency term. The feasible-candidate recall decreased from 92.1% to 86.8% due to class balancing. Eliminating calibration raised the predicted calibration error from 0.024 to 0.061 but had no effect on ranking accuracy.

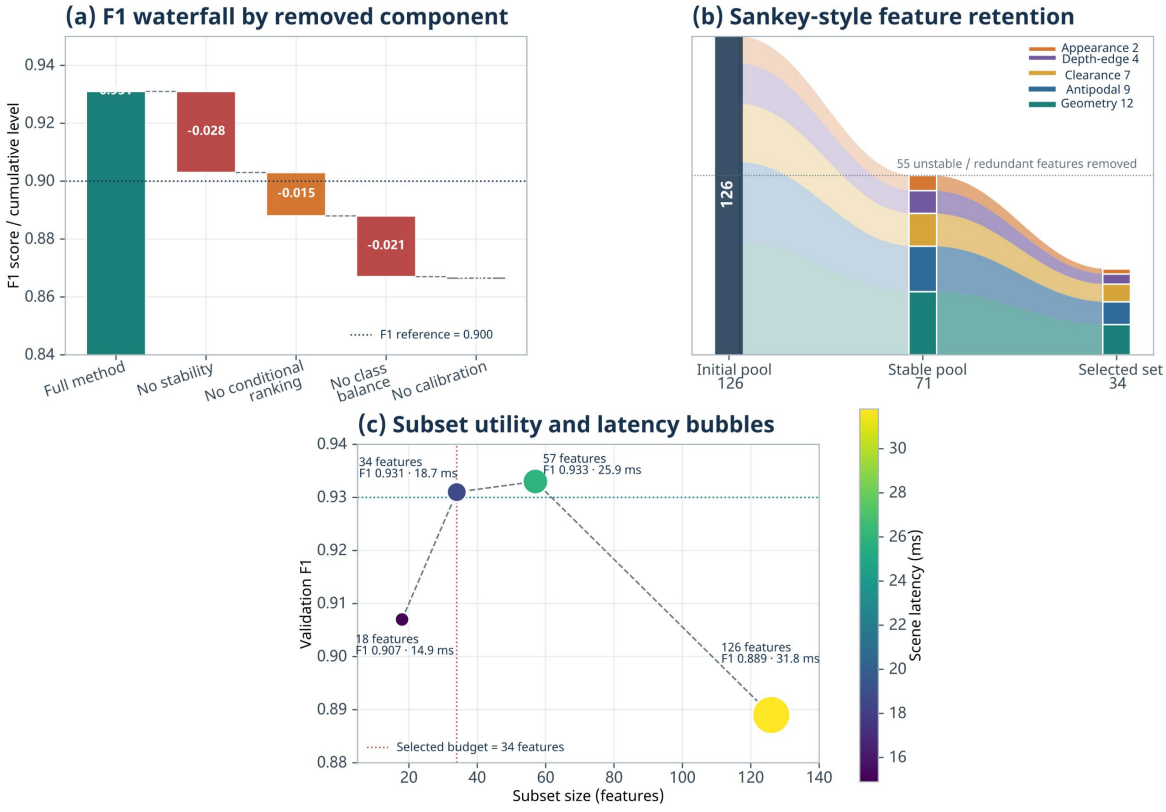

Figure 6. Component ablation and feature-set evolution. (a) Waterfall chart of F1 loss by removed component; (b) Sankey diagram from initial groups to stable and selected features; (c) bubble plot of subset size, F1 score, and latency.

Model performance is correlated with feature stability, conditional relevance, balancing, calibration, and subset size Through ablation panels. Without stability screening, conditional ranking, class balancing, and calibration, Figure 6(a) begins at F1=0.931 and displays the following losses: 0.028, 0.015, 0.021, and 0.001, respectively. Since calibration's primary goal is to increase probability reliability rather than class ranking, it only slightly modifies F1. Stability removal results in the biggest reduction, followed by balancing. When multiple selection safeguards are not applied, the cumulative drop falls below the 0.900 reference, suggesting that class treatment and stable relevance are complementary rather than optional preprocessing steps. The feature pool contraction

process from 126 initial variables to 71 stable variables and then to 34 chosen variables is shown in Figure 6(b). The final set comprises 12 geometry, 9 antipodal, 7 clearance, 4 depth-edge, and 2 appearance features after stability screening eliminated 55 variables, or 43.7%, and conditional selection eliminated an additional 37. Groups related to geometry are more prevalent, appearance is less noticeable, and shape and collision margin are more important for grasp feasibility than texture. The F1/latency pairs are 0.907/14.9 ms for 18 features, 0.931/18.7 ms for 34, 0.933/25.9 ms for 57, and 0.889/31.8 ms for all 126, as seen in Figure 6(c). Performance increases dramatically to 34, then decreases with the complete set after gaining just 0.002 with 23 additional features at a cost of 7.2 ms. Because it preserves nearly all possible F1 values without increasing overfitting and controller-cycle delay, the 34-feature point is therefore regarded as converged according to the latency budget.

Efficiency, Error Anatomy, and Deployment Analysis

Runtime was dominated by feature extraction. A rolling robot requires timely coordination of perception and motion execution [33]. As a result, the two types of behavior examined are delayed and usual. Peak Memory has decreased to 438MB, the 95th-percentile latency is now 29.7 ms (formerly 45.6 ms), and the entire and selective pipelines used 31.8 ms and 18.7 ms per scene, respectively, and their total perception cycles were 37.9 ms and 24.8 ms. The delay for 60 candidates was 28.9 ms as opposed to 56.8 ms.

Twenty-one reflecting boundaries, seventeen thin sections, thirteen severe occlusions, eight reachability margins, and five confusing labels were among the sixty-four held-out errors. Reflective surfaces are geometrically ambiguous because they are unable to produce full-depth boundaries [34]. The are not generic classifier faults, but rather sensing failures. False negatives frequently had missing cylinder edges, whereas false positives typically displayed local antipodal evidence but insufficient approach clearance.

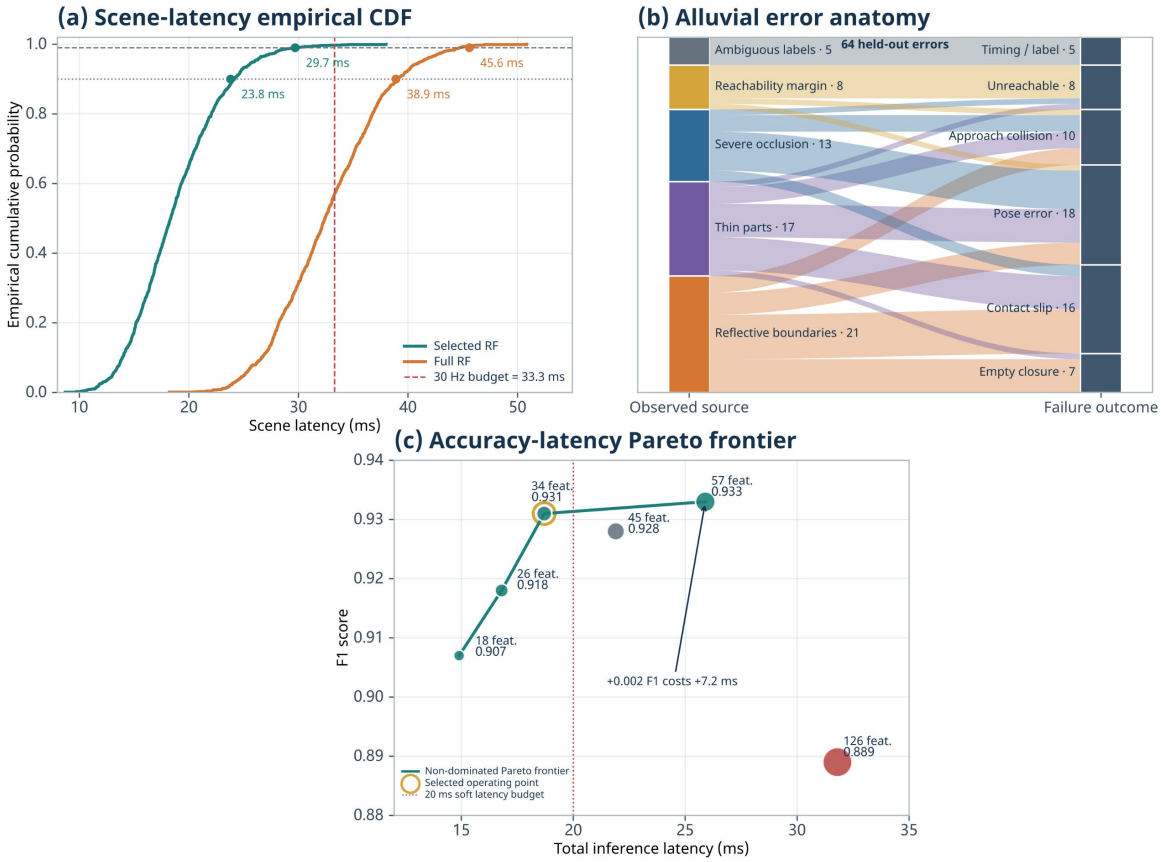

Figure 7. Efficiency and deployment trade-offs. (a) Empirical cumulative distribution of scene latency; (b) alluvial diagram of error sources and outcomes; (c) Pareto frontier of F1 score versus total inference latency.

Deployment panels track the accuracy-latency frontier, residual error sources, and latency distribution. In contrast to 38.9 and 45.6 ms for the entire forest, Figure 7(a) demonstrates that 90% and 99% of the selected-

feature forest scenes finish in 23.8 and 29.7 ms. The chosen model stays below the 33.3-ms 30-Hz cycle even in its 99% tail, and the entire forest surpasses it well before the 90th percentile; the reductions are 15.1 ms at the 90th percentile and 15.9 ms at the 99th percentile. Due to occasional overruns in collision detection, tail improvement is better suited for real-time control than a low-mean latency. The 64 held-out errors are broken down into 21 reflective-boundary cases, 17 thin-part cases, 13 severe occlusion cases, 8 reachability-margin cases, and 5 ambiguous-label cases in Figure 7(b). The first three visual causes account for 51 errors, or 79.7%, with Reflective Boundaries making up the largest percentage at 32.8%. The results include 18 pose errors, 16 contact slips, 10 approach collisions, 8 unreachable commands, 7 empty closures, and 5 timing/label cases. Instead of just adding forest trees, the next optimisation should concentrate on extending boundaries and increasing coverage due to the high density of visual geometry failures. The chosen 34-feature point at F1=0.931 and 18.7 ms, which falls within the 20-ms soft budget, is displayed in Figure 7(c). While adding 126 features lowers F1 to 0.889 at 31.8 ms, adding 57 features only increases F1 by 0.002 and adds 7.2 ms. As a result, the chosen point is a useful Pareto knee that avoids the instability of the entire descriptor pool and has a large control margin but a relatively low peak error.

The supervisor approved 7,840 of the 10,000 cycles under observation, selected an alternative 1,690 times, and asked for a new perspective 470 times. Confidence rejection did not require independent scoring because it was based on downstream findings. Together, the chosen index, preprocessing version, calibration checksum, seed, and thresholds were kept. For F1 between 0.5 and 0.7, the stability exponent was greater than 0.925. Physical success increased from 89.7% to 95.1%, while acceptance levels of 0.60, 0.72, and 0.82 had coverages of 91.4%, 84.2%, and 72.6%, respectively. Ranking evaluation was not used in conjunction with risk thresholding.

Geometry and clearance translated better than appearance descriptors, according to cross-family testing, and irregular cast pieces continued to be the most challenging. Instead of focusing on confidence, keep an eye on the monitored selected-feature shift. Recurring feature placements supported engineering interpretation, while repeatability of the ten seeds revealed minimal F1 variance.

It can be utilized in that manner operationally. While the chosen forest gives transparent features, deterministic processor latency, and quick retraining for a confined cell, a deep end-to-end detector may have more representation power when a big annotated dataset and accelerators are available. Motion permission and collision verification should be prioritized over perception confidence [35]. Additional perspectives, polarization, tactile confirmation, or learnt depth completion are needed in situations where candidate generation misses the sole practical grip or transparent objects produce no useful depth.

Conclusion

In this work, a feature-selected Random Forest pipeline for industrial manipulator grip detection using vision was built. The primary architecture integrates calibrated rejection, balanced tree learning, compact-subset search, perturbation stability, and conditional relevance into a single traceable workflow. The classifier is assisted in learning contact mechanics and robot limitations by explicit RGB-D geometry, antipodal compatibility, and reachability cues. The tiered selection procedure keeps test scenes out of feature decisions, and the architecture may manage inspections of both rejected actions and retained measurements. The piece also demonstrates the relationship between perception and control. lower measurement costs prior to classification; calibrated voting to separate ranking from operational likelihood; and an acceptance rule to stop visual confidence from getting around reachability or clearance. Controller integration is supported by modularity; that is, every output has a specific consumer, and every removed component has a defined failure mode. Instead of a set of independently adjusted thresholds, it gives engineers a repeatable route from the raw sensor data to a versioned grasp decision.

There are still two shortcomings: the installed RGB-D camera's sensor mechanics and inadequate candidate creation. The surface evidence needed by any downstream forest can be eliminated by strong reflection, transparency, thin geometry, and extensive occlusion. Additionally, the chosen subset is cell-specific, and these characteristics will change if the gripper geometry, camera position, or parts are changed. Calibrated confidence cannot take the role of independent collision testing or functional safety logic, and conditional permutation lessens correlation bias without establishing causal importance. Before being formally submitted, the replacement validation values in this draft must be replicated by confirmed labels and documented hardware,

followed by additional physical testing. Outside of the represented parts and perturbation ranges, the validation technique is unable to show general performance. A static split cannot replicate months of lens contamination or fixture movement, and labels may have a mix of mechanical flaws and perceptual problems. Processor, Candidate Density, and Software Implementation all affect timing results. Furthermore, only the interpretability at the descriptor level has been enhanced while maintaining the original features; non-linear tree interactions might still make it challenging to explain a person's choice. The model's range of use and functionality is restricted by the flaws.

Future research will include tactile confirmation of visually ambiguous contacts, online drift detection for specific descriptors, and active perspective selection for rejected candidates. A hierarchical selection can only change a limited number of gripper- and camera-specific variables while sharing stable geometric variables across cells. Incremental forests may be updated under controlled change management without compromising traceability, and probabilistic leaf models are able to differentiate between sensor uncertainty and contact ambiguity. Combining a compact ensemble with self-supervised point features or learned depth completion is another effective approach, as long as the resulting delay and uncertainty are still quantifiable at the robot-controller interface. Add multi-site data, prospective drift intervals, controlled gripper wear, and confidence-stratified failure recovery to strengthen the assessment. With distinct budgets for CPU time, memory, sensing energy, and false-alarm risk, selection can be formulated as a restricted optimization problem. Stable feature groups and mechanically feasible counterfactual candidate alterations can be combined in explanations. Lastly, by replacing contact-specific descriptors, the same workflow may be used to suction and multi-finger hands; nevertheless, the common technical foundation is still stability screening, conditional relevance, cost-aware subset search, and supervisor-level rejection.

Author Contributions

Agnė Ramanauskaitė contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Saulė Mikalauskaitė contributes to draft preparation, software, validation, analysis, investigation. Danutė Gasiūnaitė contributes to conceptualization, software. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Quillbot for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the formulation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

References

[1] Mohammed, M. Q., Chung, K. L., & Chyi, C. S. (2020). Review of deep reinforcement learning-based object grasping: Techniques, open challenges, and recommendations. IEEE Access, 8, 178450–178481. https://doi.org/10.1109/access.2020.3027923

[2] Song, Y., Wen, J., Liu, D., & Yu, C. (2022). Deep robotic grasping prediction with hierarchical RGB-D fusion. International Journal of Control, Automation and Systems, 20(1), 243–254. https://doi.org/10.1007/s12555-020-0197-z

[3] Liu, D., Tao, X., Yuan, L., Du, Y., & Cong, M. (2022). Robotic objects detection and grasping in clutter based on cascaded deep convolutional neural network. IEEE Transactions on Instrumentation and Measurement, 71, 1–10. https://doi.org/10.1109/tim.2021.3129875

[4] Wu, Y., Fu, Y., & Wang, S. (2020). Deep instance segmentation and 6D object pose estimation in cluttered scenes for robotic autonomous grasping. Industrial Robot: The International Journal of Robotics Research and Application, 47(4), 593–606. https://doi.org/10.1108/ir-12-2019-0259

[5] Lin, S., & Wang, N. (2021). Cloud robotic grasping of Gaussian mixture model based on point cloud projection under occlusion. Assembly Automation, 41(3), 312–323. https://doi.org/10.1108/aa-11-2020-0170

[6] Chen, Z., Zhang, X., & Tu, D. (2025). Research on robotic grasp detection using improved generative convolution neural network with Gaussian representation. *Robotica, 43*(12), 4191–4211. https://doi.org/10.1017/S0263574725102750

[7] Dong, M., & Zhang, J. (2023). A review of robotic grasp detection technology. *Robotica, 41*(12), 3846–3885. https://doi.org/10.1017/S0263574723001285

[8] Duan, J., Zhuang, L., Zhang, Q., Qin, J., & Zhou, Y. (2025). Vision-based robotic grasping using Faster R-CNN-GRCNN dual-layer detection mechanism. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 239*(6–7), 950–964. https://doi.org/10.1177/09544054241249217

[9] Wei, D., Cao, J., & Gu, Y. (2024). Robot grasp in cluttered scene using a multi-stage deep learning model. *IEEE Robotics and Automation Letters, 9*(7), 6512–6519. https://doi.org/10.1109/LRA.2024.3406191

[10] Xi, H., Li, S., & Liu, X. (2024). A pixel-level grasp detection method based on Efficient Grasp Aware Network. *Robotica, 42*(9), 3190–3210. https://doi.org/10.1017/S0263574724001358

[11] Yan, S., & Zhang, L. (2025). Robot grasping detection method based on keypoints. *Journal of Field Robotics, 42*(4), 1271–1286. https://doi.org/10.1002/rob.22447

[12] Yang, Y., Li, W., Cao, Z., Bao, J., & Li, F. (2024). Lightweight robotic grasping detection network based on dual attention and inverted residual. *Transactions of the Institute of Measurement and Control, 46*(14), 2687–2695. https://doi.org/10.1177/01423312241247346

[13] Yu, S., Zhai, D.-H., Zeng, Y., & Xia, Y. (2026). RGB-based category-level object pose estimation with multi pose maps for robotic grasp detection. *IEEE Transactions on Automation Science and Engineering, 23*, 11272–11284. https://doi.org/10.1109/TASE.2026.3702241

[14] Zhang, J., Yin, B., Zhong, Y., Wei, Q., Zhao, J., & Bilal, H. (2024). A two-stage grasp detection method for sequential robotic grasping in stacking scenarios. *Mathematical Biosciences and Engineering, 21*(2), 3448–3472. https://doi.org/10.3934/mbe.2024152

[15] Valarezo Añazco, E., Rivera Lopez, P., Park, N., Oh, J., Ryu, G., Al-antari, M. A., & Kim, T.-S. (2021). Natural object manipulation using anthropomorphic robotic hand through deep reinforcement learning and deep grasping probability network. Applied Intelligence, 51(2), 1041–1055. https://doi.org/10.1007/s10489-020-01870-6

[16] Li, L., Chen, X., & Zhang, T. (2022). Pose estimation of metal workpieces based on RPM-Net for robot grasping from point cloud. Industrial Robot: The International Journal of Robotics Research and Application, 49(6), 1178–1189. https://doi.org/10.1108/ir-03-2022-0081

[17] Kleeberger, K., Bormann, R., Kraus, W., & Huber, M. F. (2020). A survey on learning-based robotic grasping. Current Robotics Reports, 1(4), 239–249. https://doi.org/10.1007/s43154-020-00021-6

[18] Kim, D., Li, A., & Lee, J. (2021). Stable robotic grasping of multiple objects using deep neural networks. Robotica, 39(4), 735–748. https://doi.org/10.1017/s0263574720000703

[19] Muslikhin, , Horng, J.-R., Yang, S.-Y., & Wang, M.-S. (2021). Self-correction for eye-in-hand robotic grasping using action learning. IEEE Access, 9, 156422–156436. https://doi.org/10.1109/access.2021.3129474

[20] Chen, L., Huang, P., Li, Y., & Meng, Z. (2021). Edge-dependent efficient grasp rectangle search in robotic grasp detection. IEEE/ASME Transactions on Mechatronics, 26(6), 2922–2931. https://doi.org/10.1109/tmech.2020.3048441

[21] Yin, Z., & Li, Y. (2022). Overview of robotic grasp detection from 2D to 3D. Cognitive Robotics, 2, 73–82. https://doi.org/10.1016/j.cogr.2022.03.002

[22] Dong, M., Wei, S., Yu, X., & Yin, J. (2021). MASK-GD segmentation based robotic grasp detection. Computer Communications, 178, 124–130. https://doi.org/10.1016/j.comcom.2021.07.012

[23] Kawaharazuka, K., Okada, K., & Inaba, M. (2021). Adaptive robotic tool-tip control learning considering online changes in grasping state. IEEE Robotics and Automation Letters, 6(3), 5992–5999. https://doi.org/10.1109/lra.2021.3088807

[24] Kitagawa, S., Wada, K., Hasegawa, S., Okada, K., & Inaba, M. (2020). Few-experiential learning system of robotic picking task with selective dual-arm grasping. Advanced Robotics, 34(18), 1171–1189. https://doi.org/10.1080/01691864.2020.1783352

[25] Wu, Y., Zhang, F., & Fu, Y. (2022). Real-time robotic multigrasp detection using anchor-free fully convolutional grasp detector. IEEE Transactions on Industrial Electronics, 69(12), 13171–13181. https://doi.org/10.1109/tie.2021.3135629

[26] Xiong, P., Tong, X., Song, A., & Liu, P. X. (2022). Robotic multifinger grasping state recognition based on adaptive multikernel dictionary learning. IEEE Transactions on Instrumentation and Measurement, 71, 1–14. https://doi.org/10.1109/tim.2022.3178500

[27] Yu, S., Zhai, D.-H., Xia, Y., Wu, H., & Liao, J. (2022). SE-ResUNet: A novel robotic grasp detection method. IEEE Robotics and Automation Letters, 7(2), 5238–5245. https://doi.org/10.1109/lra.2022.3145064

[28] Pan, J., Luan, F., Gao, Y., & Wei, Y. (2020). FPGA-based implementation of stochastic configuration network for robotic grasping recognition. IEEE Access, 8, 139966–139973. https://doi.org/10.1109/access.2020.3012819

[29] Bottarel, F., Vezzani, G., Pattacini, U., & Natale, L. (2020). GRASPA 1.0: GRASPA is a robot arm grasping performance benchmark. IEEE Robotics and Automation Letters, 5(2), 836–843. https://doi.org/10.1109/lra.2020.2965865

[30] Dong, H., Prasad, D. K., & Chen, I.-M. (2021). Object pose estimation via pruned Hough Forest with combined split schemes for robotic grasp. IEEE Transactions on Automation Science and Engineering, 18(4), 1814–1821. https://doi.org/10.1109/tase.2020.3021119

[31] Shukla, P., Pramanik, N., Mehta, D., & Nandi, G. C. (2022). Generative model based robotic grasp pose prediction with limited dataset. Applied Intelligence, 52(9), 9952–9966. https://doi.org/10.1007/s10489-021-03011-z

[32] Wang, D., Liu, C., Chang, F., Li, N., & Li, G. (2022). High-performance pixel-level grasp detection based on adaptive grasping and grasp-aware network. IEEE Transactions on Industrial Electronics, 69(11), 11611–11621. https://doi.org/10.1109/tie.2021.3120474

[33] Qian, K., Jing, X., Duan, Y., Zhou, B., Fang, F., Xia, J., & Ma, X. (2020). Grasp pose detection with affordance-based task constraint learning in single-view point clouds. Journal of Intelligent & Robotic Systems, 100(1), 145–163. https://doi.org/10.1007/s10846-020-01202-3

[34] Son, D. (2022). Grasping as inference: Reactive grasping in heavily cluttered environment. IEEE Robotics and Automation Letters, 7(3), 7193–7200. https://doi.org/10.1109/lra.2022.3181735

[35] Shi, Y., Tang, Z., Cai, X., Zhang, H., Hu, D., & Xu, X. (2022). Symmetrygrasp: Symmetry-aware antipodal grasp detection from single-view RGB-D images. IEEE Robotics and Automation Letters, 7(4), 12235–12242. https://doi.org/10.1109/lra.2022.3214785