applied sciences

MDPI

Article

From Model Proliferation to Deployment Intelligence: A Risk-Traceable System-Aware Model Selection Framework

Vishakha Singh and Phisan Kaewprapha *

Department of Electrical and Computer Engineering, Faculty of Engineering, Thammasat School of Engineering, Thammasat University, Pathum Thani 12120, Thailand; vish12150@gmail.com
* Correspondence: kphisan@engr.tu.ac.th

Abstract

The rapid growth of machine-learning architectures has shifted an important engineering challenge from model development to selecting the most suitable trained candidate for a specific deployment context. This study proposes the System-Aware Model Selection (SMS) Framework, implemented through the Deployment-Aware Model Selection Algorithm (DAMSA), as a three-layer post-training decision architecture integrating risk-to-criterion mapping, candidate control, reference ranking, and robustness analysis. Across five engineering scenarios, DAMSA-SMS was compared with benchmark-oriented and conventional decision procedures under common candidate evidence and priority assumptions. Conventional multi-criteria procedures generally agreed on the nominal reference winner, but DAMSA-SMS revealed different levels of decision robustness. YOLOv7 was selected for traffic monitoring and pharmaceutical inspection with first-rank acceptabilities of 92.8% and 91.2%, respectively. In defense surveillance, YOLOv8s was preferred to the highest-mAP YOLOv11s because its Precision was 18.9 percentage points higher. YOLOv8s was Pareto-unique and preference-stable in microplastics detection. In CCPP monitoring, Random Forest was preferred under the Testing-Generalization profile, with the decision switching below a 12.99% Generalization Stability weight. Removing F1-score did not change any vision recommendation. These results show that nominal ranking agreement does not necessarily imply deployment robustness.

Keywords: machine-learning model selection; deployment intelligence; operational risk; Pareto screening; preference robustness; first-rank acceptability; benchmark evaluation; multi-criteria decision making

Check for updates

Academic Editors: Georgia Dede and Thomas Kamalakis

Received: 1 September 2026
Revised: 15 October 2026
Accepted: 19 October 2026
Published: 22 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

1.1. Research Background and Existing Challenges

Artificial intelligence has become central to modern engineering systems, supporting applications in computer vision, transportation, manufacturing, healthcare, environmental monitoring, energy, and defense. Advances in machine learning, deep neural networks, optimization methods, and computing hardware have substantially increased the number of available models and trained variants [1–3]. Consequently, an important engineering challenge arises after model development: selecting which trained candidate should be deployed in a particular operational environment.

Training and benchmarking establish how candidate models perform, whereas deployment selection must interpret those results according to system objectives, operational risks, constraints, and stakeholder priorities. These requirements can differ substantially

across applications. For example, in traffic monitoring, missed detections may affect counting, tracking, and congestion analysis; in defense surveillance, false-positive detections may generate unnecessary alerts; in pharmaceutical inspection, missed defects and false rejection decisions have different quality consequences; in microplastics detection, incomplete detection may affect environmental analysis; and in power-plant monitoring, testing performance may need to be considered together with train-test consistency. Thus, the operational significance of model errors depends on the scenario.

Benchmark metrics provide essential and reproducible evidence of model behavior. Object-detection systems are commonly evaluated using mean Average Precision (mAP), Precision, Recall, and F1-score, while regression and classification systems use measures such as accuracy, coefficient of determination, mean absolute error, and root mean square error [4–6]. However, benchmark results alone do not determine the operational importance of competing performance characteristics. A model that performs best on one metric may perform less well on the criterion most closely associated with deployment risk.

1.2. Existing Methods, Limitations, and Research Gap

Existing approaches address different parts of this problem. Model Cards, robustness studies, and AI risk-management frameworks emphasize intended use, evaluation context, limitations, and operational risk [7–9], while Green AI research highlights latency, memory, energy consumption, model size, and computational cost as potential deployment considerations [10–12]. Multi-criteria decision-making (MCDM) methods such as AHP, SAW, TOPSIS, VIKOR, PROMETHEE, and ELECTRE provide established mechanisms for priority elicitation and candidate ranking [13–15]. However, these methods generally operate after the alternatives, criteria, decision matrix, and preference structure have been defined.

The main gap addressed in this study is not the lack of ranking methods, but the lack of an integrated process around deployment-oriented model selection. Before models are ranked, operational objectives and risks need to be translated into measurable criteria, with a clear distinction between direct, derived, descriptive, and mandatory evidence. Infeasible or Pareto-dominated candidates should then be screened out before preference-based ranking. Finally, the resulting recommendation should be tested to determine whether it remains stable when a derived criterion is removed, deployment priorities change, or small variations occur in the reported benchmark results.

This gap is therefore broader than the choice of a ranking operator alone. Conventional multi-criteria decision-making procedures can provide effective mechanisms for weighting, aggregation, compromise selection, or outranking once the alternatives, criteria, measurements, and preference structure have already been specified. The unresolved deployment-selection problem considered here also includes how operational risks are translated into measurable criteria before ranking, how different types of evidence are assigned appropriate decision roles, how candidate eligibility is controlled, and how the resulting recommendation is examined afterwards for sensitivity to alternative priorities and benchmark variation.

1.3. Proposed Approach, Research Questions, and Objectives

To address this gap, this study proposes the System-Aware Model Selection (SMS) Framework, implemented through the Deployment-Aware Model Selection Algorithm (DAMSA), as a three-layer approach for post-training model selection. The first layer translates deployment objectives and risks into measurable criteria and defines the role of each criterion. The second layer screens candidates using feasibility and Pareto analysis before applying the transparent SMS Index for reference ranking. The final layer examines

whether the recommendation remains stable when composite criteria are removed, deployment priorities change, or small variations occur in the reported benchmark results. The SMS Index is used as the transparent reference-ranking mechanism within this architecture; it is not proposed as a mathematically superior replacement for established MCDM ranking operators.

Based on the identified gap, this study addresses the following research questions:

RQ1. How can deployment objectives, operational risks, and measured constraints be systematically translated into measurable criteria for post-training model selection?
RQ2. How can criterion roles and candidate eligibility be controlled before preference-based ranking so that directly measured, derived, descriptive, and mandatory evidence are treated appropriately?
RQ3. When conventional decision procedures agree on a reference-ranked model, does this agreement imply that the deployment recommendation is robust to admissible changes in deployment priorities and benchmark evidence?
RQ4. How can preference sensitivity and performance-measurement sensitivity be quantified and interpreted across heterogeneous engineering deployment scenarios?

Accordingly, the objectives of this study are to:

1. Develop a risk-traceable post-training model-selection architecture that translates deployment objectives, operational risks, and measured constraints into measurable decision criteria.
2. Establish criterion-role classification, feasibility control, and Pareto screening before preference-based ranking.
3. Compare the reference recommendations produced under DAMSA-SMS with benchmark-oriented and conventional decision procedures under common candidate evidence and priority assumptions.
4. Evaluate whether the resulting recommendations remain stable under composite-metric removal, admissible priority variation, and bounded perturbation of benchmark measurements.

The principal contributions of this study are fourfold:

1. Risk-traceable decision construction. DAMSA-SMS translates deployment objectives, operational risks, and measured constraints into a scenario-specific decision model before candidate ranking.
2. Criterion-role and candidate control. Directly measured core criteria, derived composite criteria, descriptive measurements, and mandatory requirements are assigned distinct roles. Feasibility and Pareto screening are then used to restrict formal ranking to eligible and non-dominated candidates.
3. Comparative decision evaluation. DAMSA-SMS is evaluated alongside benchmark-oriented and conventional decision procedures under common candidate evidence and reference-priority assumptions. This comparison examines whether agreement at a fixed reference operating point is sufficient to establish a robust deployment recommendation.
4. Post-ranking robustness analysis. The framework evaluates whether the recommendation depends on a composite criterion, how the preferred candidate changes across admissible deployment priorities, and whether bounded variation in the reported benchmark measurements can alter the final selection.

The methodology is evaluated across five engineering scenarios: smart traffic monitoring, defense vehicle surveillance, pharmaceutical tablet inspection, microplastics detection, and Combined Cycle Power Plant (CCPP) monitoring. The scenarios were selected to examine procedural transferability across heterogeneous decision conditions, including

object-detection and regression tasks, different dataset scales, different operational hazards, different criterion-priority structures, and cases ranging from a single Pareto-dominant candidate to multiple candidates involving substantive performance trade-offs. Rather than producing only a model ranking, the framework provides a traceable recommendation that explains why a model is selected and when another candidate may become a better choice.

1.4. Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews related work and positions the research gap. Sections 3 and 4 present the SMS Framework and formalize DAMSA and the SMS Index. Section 5 describes the experimental validation protocol. Section 6 reports the results, and Section 7 discusses their implications and limitations. Section 8 concludes the study.

2. Literature Review

2.1. Model Proliferation and Post-Training Selection

Advances in deep learning, transformers, and real-time object detection have greatly increased the number of machine-learning models available for practical applications [16–20]. Most conventional model-selection approaches focus on the development stage through methods such as validation, hyperparameter optimization, neural architecture search, or automated machine learning. In contrast, this study focuses on a later stage: choosing among models that have already been trained and benchmarked. At this stage, the preferred model may change depending on deployment priorities. For example, an application that prioritizes Recall may favor a different model from one focused on reducing false positives or maintaining train-test consistency. Post-training selection therefore requires benchmark results to be considered together with the needs of the intended deployment environment.

2.2. Benchmark Evaluation and Deployment Context

Benchmark evaluation provides reproducible evidence for comparing candidate models. In object detection, established practices include mAP, Precision, Recall, and F1-score, while regression applications commonly use measures such as coefficient of determination, mean absolute error, and root mean square error [4–6].

However, benchmark metrics do not independently determine the operational importance of performance differences. A Recall improvement may be critical when missed detections create the primary risk, whereas Precision may be more important when false alarms have greater operational consequences. Robustness research, Model Cards, and the NIST AI Risk Management Framework further emphasize evaluation conditions, intended use, limitations, and operational context [7–9]. These approaches strengthen model evaluation and risk awareness, but do not by themselves provide a complete procedure for selecting one trained candidate when deployment priorities differ.

2.3. Sustainable AI and Resource-Aware Evaluation

Green AI research has highlighted the computational, financial, energy, and environmental costs of machine learning [10–12]. Resource-related factors such as latency, memory consumption, model size, computational demand, and energy use can therefore be relevant to deployment decisions [21]. Such criteria should, however, be incorporated only when comparable measurements are available across candidates. DAMSA-SMS consequently allows resource measures to be represented as active criteria or hard constraints, but does not include them in the present validation because consistent measurements were un-

available for all candidate models. Accordingly, this study does not claim experimentally demonstrated improvements in computational efficiency or environmental impact.

2.4. Multi-Criteria Decision-Making for Model Selection

Multi-criteria decision-making (MCDM) provides established methods for evaluating alternatives under competing criteria. AHP supports priority elicitation through pairwise comparison; SAW uses weighted aggregation; TOPSIS evaluates proximity to an ideal solution; and VIKOR, PROMETHEE, and ELECTRE provide compromise or outranking mechanisms [13–15].

These methods are applicable once alternatives, criteria, measurements, criterion directions, and preference information have been established. However, they do not intrinsically determine how deployment hazards should be translated into measurable model criteria, how direct and derived evidence should be distinguished, or how the resulting recommendation should subsequently be tested for sensitivity to deployment priorities and benchmark variation. DAMSA-SMS addresses this surrounding decision architecture rather than proposing the SMS Index as a mathematically superior replacement for established MCDM methods.

2.5. Pareto Efficiency, Criterion Roles, and Preference Uncertainty

Pareto analysis identifies candidates that are consistently weaker across the core criteria without requiring criterion weights [22]. This helps reduce the candidate set before ranking while retaining models that involve meaningful performance trade-offs.

Criterion dependence is also relevant when constructing the decision model. For example, F1-score is derived from Precision and Recall and therefore represents their balance rather than an independent measurement [5]. In addition, deployment priorities may not be represented adequately by a single fixed weight vector. Rank-order approaches can derive weights from stated priority structures [23], while Stochastic Multicriteria Acceptability Analysis (SMAA) evaluates candidate preferences across alternative weight combinations [24]. These concepts motivate the criterion-role control and first-rank acceptability analysis used within DAMSA-SMS.

2.6. Research Gap and Positioning

Existing literature provides complementary components of deployment-oriented model evaluation. Benchmarking provides reproducible performance evidence [4–6]; Model Cards and AI risk-management frameworks emphasize intended use and operational context [8,9]; Green AI highlights resource efficiency [10–12,21]; MCDM methods support priority elicitation and candidate ranking [13–15]; Pareto analysis identifies efficient alternatives [22]; and SMAA-type methods examine preference uncertainty [24].

However, these components are generally applied separately. A complete post-training deployment decision also requires operational objectives and risks to be translated into measurable criteria before ranking, criterion roles and mandatory requirements to be controlled during candidate evaluation, and the resulting recommendation to be examined afterward for dependence on composite criteria, alternative priority profiles, and variation in benchmark measurements.

DAMSA-SMS addresses this integration gap through a three-layer architecture linking decision construction, candidate control and reference ranking, and post-ranking robustness analysis. Its contribution therefore concerns the organization and traceability of the deployment-selection process rather than mathematical superiority over individual benchmark or MCDM techniques.

More specifically, DAMSA-SMS differs from conventional single-stage ranking procedures in the scope of the decision process that it formalizes. Methods such as SAW, TOPSIS,

VIKOR, PROMETHEE, ELECTRE, and related MCDM approaches primarily determine how alternatives are ranked once the candidate set, criteria, measurements, and preference information have been established. Pareto analysis identifies non-dominated alternatives without specifying a unique preferred candidate, whereas SMAA-type procedures examine ranking acceptability under preference uncertainty. DAMSA-SMS instead connects these decision components within a deployment-oriented workflow: operational objectives and risks are translated into measurable criteria before ranking; criterion roles and candidate eligibility are controlled; a transparent reference ranking is produced; and the resulting recommendation is subsequently examined for dependence on composite criteria, admissible priority variation, and bounded benchmark variation.

Accordingly, the SMS Index should not be interpreted as a new general-purpose MCDM operator or as a mathematically superior alternative to established ranking methods. Its role is to provide an interpretable reference ranking within the wider DAMSA-SMS architecture. The methodological contribution lies primarily in the risk traceability, criterion-role control, candidate screening, and robustness analyses surrounding that reference ranking.

Table 1 positions DAMSA-SMS relative to representative benchmark and model-selection approaches. The comparison is intended as a functional positioning analysis rather than evidence of numerical superiority. Conventional MCDM procedures remain applicable for preference elicitation or final ranking once the decision structure has been established, while Pareto and SMAA-type methods provide complementary candidate-screening and preference-uncertainty functions. DAMSA-SMS integrates these stages with risk-to-criterion construction and post-ranking robustness analysis. Section 5 therefore supplements this functional comparison with controlled comparative decision procedures using common candidate evidence and reference-priority assumptions.

Table 1. Functional comparison of deployment-oriented decision capabilities.

Approach	Risk Traceability	Criterion Roles	Pareto Screening	Composite Ablation	Policy Robustness	Metric Perturbation
Benchmark-only	No	No	No	No	No	No
AHP	Not intrinsic	Not intrinsic	Not intrinsic	Not intrinsic	Limited externally	Not intrinsic
SAW	Not intrinsic	Not intrinsic	Not intrinsic	Not intrinsic	Limited externally	Not intrinsic
TOPSIS	Not intrinsic	Not intrinsic	Not intrinsic	Not intrinsic	Limited externally	Not intrinsic
Pareto + SAW	Not intrinsic	Not intrinsic	Yes	Not intrinsic	Limited externally	Not intrinsic
SMAA/rank acceptability	Not intrinsic	Not intrinsic	Not intrinsic	Not intrinsic	Yes	Not intrinsic
DAMSA-SMS	Yes	Yes	Yes	Yes	Yes	Yes

3. System-Aware Model Selection Framework

3.1. Framework Purpose and Components

The System-Aware Model Selection (SMS) Framework supports post-training selection among candidate machine-learning models that have already been trained and benchmarked. Rather than modifying model architectures, training procedures, hyperparameters, or learned parameters, it interprets the available model evidence according to

the objectives, operational risks, constraints, and priorities of the intended deployment environment. The objective is therefore not to identify a universally superior model, but to produce a traceable recommendation for a documented deployment context.

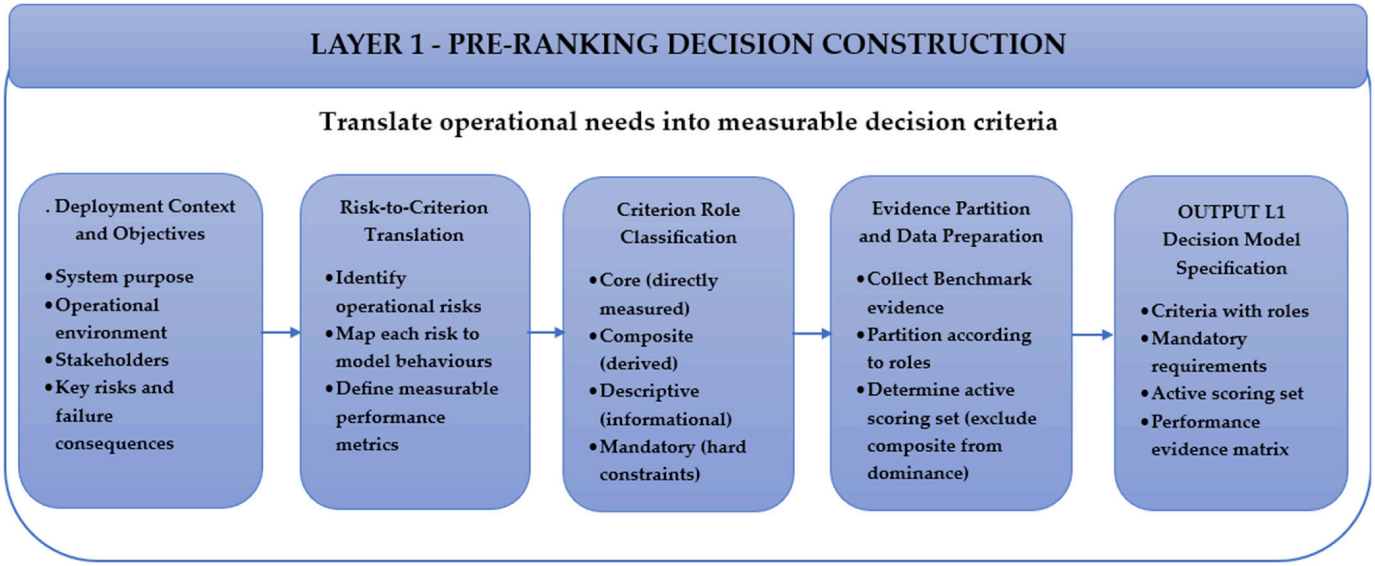
The proposed approach comprises three connected components: the SMS Framework, the Deployment-Aware Model Selection Algorithm (DAMSA), and the SMS Index. The SMS Framework defines the overall deployment-decision architecture; DAMSA implements the sequence of decision construction, candidate control, ranking, and robustness analysis; and the SMS Index provides the transparent reference-scoring mechanism after the deployment-specific decision structure has been established.

3.2. Framework Architecture

The framework is organized into three connected layers, as shown in Figure 1:

1. Pre-ranking decision construction: deployment objectives, operational risks, and measured constraints are translated into measurable criteria, which are assigned core, composite, mandatory, or descriptive roles.
2. Candidate control and reference ranking: measured hard constraints are applied where available, Pareto screening identifies efficient candidates using the core criteria, and the SMS Index produces the reference ranking.
3. Post-ranking robustness analysis: the recommendation is examined through composite-metric ablation, preference-robustness analysis, and bounded performance-metric perturbation to determine whether the selected model remains defensible under alternative decision conditions.

The separation of these layers distinguishes how the deployment decision is constructed, how candidates are controlled and ranked, and how the resulting recommendation is stress-tested. The SMS Index therefore serves as the reference aggregation mechanism within the broader DAMSA-SMS architecture rather than constituting the complete methodology itself.



(a)

Figure 1. *Cont.*

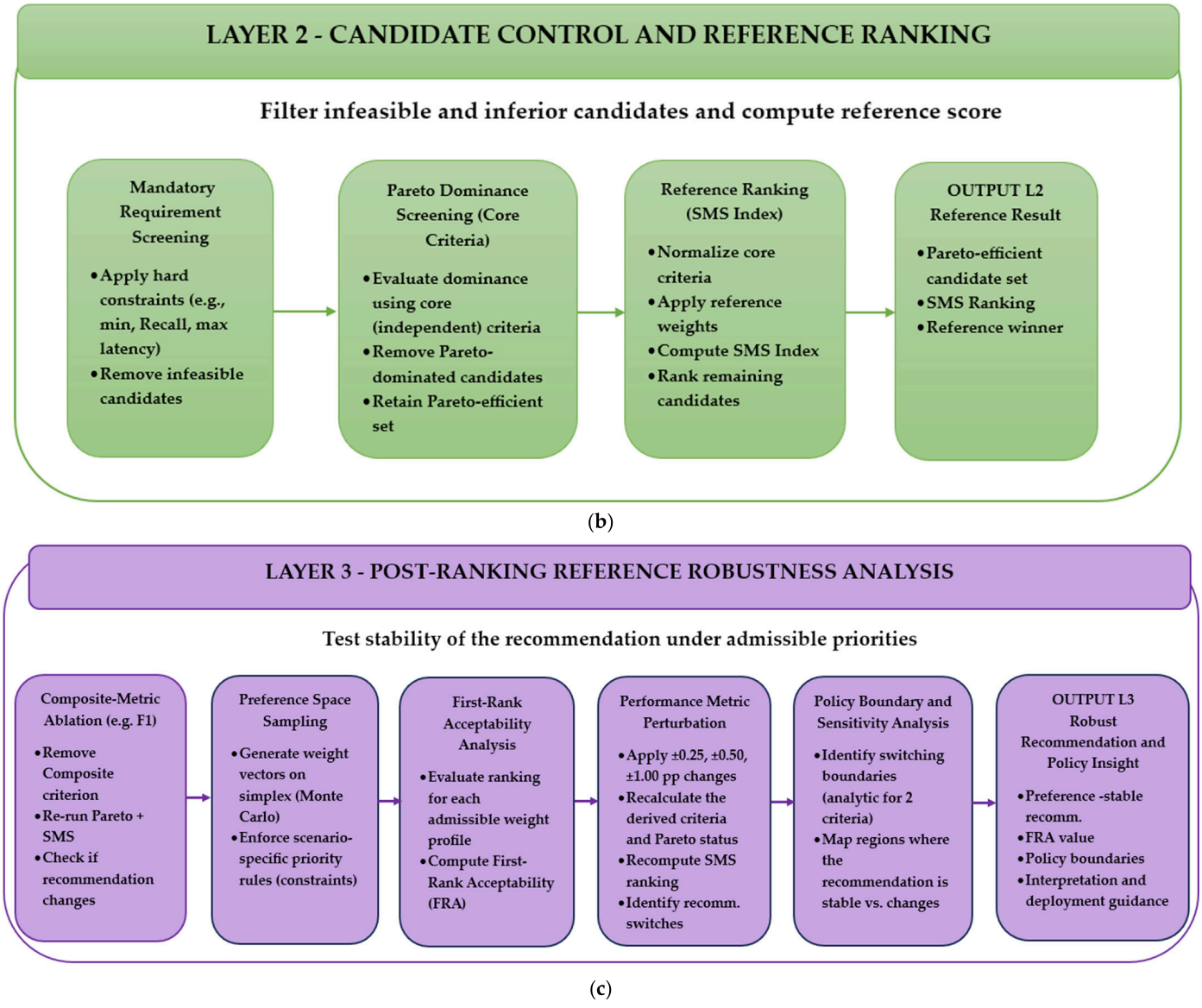


Figure 1. Three-layer DAMSA-SMS architecture and decision flow: (**a**) deployment objectives, risks, constraints, and candidate evidence are translated into role-classified criteria; (**b**) candidate eligibility is controlled through hard constraints and Pareto screening before reference SMS ranking; and (**c**) the resulting recommendation is stress-tested through composite-metric ablation, preference variation, and bounded metric perturbation.

The three layers operate sequentially and exchange explicit decision outputs. The pre-ranking layer receives the deployment specification ($D = (O,R,C)$) together with the available candidate evidence and transforms these inputs into role-classified decision criteria, including core, composite, mandatory, and descriptive evidence. These outputs form the input to the candidate-control and reference-ranking layer, where hard constraints are applied, Pareto-efficient candidates are identified, and the reference priority profile is used to compute the SMS ranking. The resulting Pareto set, reference winner, score margins, and criterion structure are then passed to the post-ranking layer. This final layer does not modify the trained models but instead re-evaluates the decision under composite-criterion removal, admissible priority variation, and bounded metric perturbation. The final output is therefore not only a ranked candidate, but a recommendation trace containing the selected model, conditional alternatives where applicable, and evidence describing the sensitivity of the decision to both policy assumptions and benchmark variation.

3.3. Framework Outputs and Scope

For each deployment scenario, DAMSA-SMS produces a risk-to-criterion mapping, classified criterion set, feasible and Pareto-efficient candidate set, reference SMS ranking, composite-metric ablation result where applicable, preference-robustness result, metric-perturbation result, and final recommendation or conditional alternative. The framework is general at the procedural level but scenario-specific at the decision level. Criteria such as latency, memory consumption, model size, computational demand, or energy use may be incorporated when comparable measurements are available across candidates. In the present study, validation is restricted to the predictive-performance and train–test consistency measurements available in the experimental decision matrices; therefore, no experimental claim is made regarding improvements in computational efficiency, energy consumption, or emissions.

The applicability of DAMSA-SMS is not restricted to a particular data dimensionality or machine-learning model family. Its prerequisites are that: (i) a finite set of trained candidate models is available; (ii) the evaluation criteria are measurable and comparable across candidates; (iii) the preference direction of each active criterion is defined; (iv) deployment relevance can be documented for the selected criteria; and (v) any hard constraints are measurable consistently across the candidate set. Criteria may therefore differ across scenarios; for example, object-detection metrics such as mAP, Precision, and Recall or regression-oriented measures such as Testing Performance and Generalization Stability may be used, provided that they are transformed into a common suitability direction before aggregation. The framework is thus general at the procedural level while remaining scenario-specific in its criterion definitions and deployment priorities.

4. DAMSA Methodology and SMS Index

This section formalizes the three DAMSA-SMS layers introduced in Section 3. The methodology constructs the deployment-specific decision model, controls candidate eligibility through feasibility and Pareto screening, produces a reference ranking using the SMS Index, and evaluates recommendation robustness through F1-score ablation, preference variation, and bounded metric perturbation. Table 2 below gives the main notation used in DAMSA-SMS.

Table 2. Main notation used in DAMSA-SMS.

Symbol	Definition	Range/Role
(D)	Deployment scenario	$(D = (O,R,C))$
(O)	Operational objectives	Scenario-specific
(R)	Operational risks	Scenario-specific
(C)	Measured deployment constraints	Scenario-specific
$\mathcal{M}$	Set of trained candidate models	Finite set
(M_i)	Candidate model (i)	$(i = 1, \ldots, m)$
(E_{core})	Directly measured core criteria	Used in Pareto screening and scoring
(E_{comp})	Derived composite criteria	Used in scoring where applicable
(H)	Hard constraints	Applied before ranking
(E_{desc})	Descriptive measurements	Not used directly in ranking
(E_{active})	Active scoring criteria	$(E_{core} \cup E_{comp})$

Table 2. *Cont.*

Symbol	Definition	Range/Role
(x_{ij})	Observed value of candidate (i) on criterion (j)	Criterion-specific
$(\hat{x}_{ij})$	Normalized criterion value	([0,1])
(w_j)	Criterion weight	$(w_j \geq 0, \sum_j w_j = 1)$
$(\mathcal{F}(D))$	Feasible candidate set	*Subset of* $(\mathcal{M})$
$(\mathcal{P}(D))$	Pareto-efficient candidate set	*Subset of* $(\mathcal{F}(D))$
$(\mathcal{W}(D))$	Admissible priority region	Constrained simplex
(B)	Number of accepted priority samples	$(B = 500{,}000)$ for vision scenarios
$(\hat{a}_i^1)$	Estimated first-rank acceptability	([0,1])
(δ)	Metric-perturbation magnitude	({0.25,0.50,1.00 }) *pp*

4.1. Problem Formulation

Let the finite set of trained candidate models available for a deployment scenario be $\mathcal{M} = \{M^1, M^2, \ldots, M_m\}$. Then the deployment scenario is represented as:

$$D = (O, R, C) \quad (1)$$

where O denotes operational objectives, R denotes operational risks, and C denotes measured deployment constraints.

The operational-to-empirical translation function partitions the evidence into formal decision roles:

$$\Phi(D) = (E_{core}, E_{comp}, H, E_{desc}) \quad (2)$$

where E_{core} contains directly measured criteria, E_{comp} contains derived composite criteria, H contains hard constraints, and E_{desc} contains descriptive measurements.

The mapping (Φ) is not a learned transformation, statistical estimator, or optimization function. It represents the documented engineering reasoning used to translate a deployment specification into measurable decision evidence before ranking. The mapping is scenario-specific: an operational risk associated with missed detections may motivate Recall as a core criterion; a false-alert burden may motivate Precision; a requirement for reliable object localization may motivate mAP; and concern about train–test consistency may motivate Generalization Stability. The purpose of (Φ) is therefore to make the link between deployment need and evaluation criterion explicit and traceable rather than to infer criteria automatically from data. An illustrative operational-risk-to-criterion mapping is provided in Table 3.

Unless otherwise stated, all criteria used for formal ranking are required to be measurable for every candidate in the comparison set and to have an explicitly defined preference direction. Percentage-based benefit criteria are treated as larger-is-better quantities, whereas future cost-type criteria must be transformed into corresponding suitability measures before entering the SMS calculation.

The active scoring set is:

$$E_{active} = E_{core} \cup E_{comp} \quad (3)$$

Only criteria that are operationally relevant and consistently measured across candidate models are retained as active criteria.

Table 3. Illustrative operational-risk-to-criterion mapping used in DAMSA-SMS.

Deployment Context	Operational Concern	Engineering Interpretation	Criterion
Traffic monitoring	Missed vehicles	Incomplete detection affects counting/tracking	Recall
Defense surveillance	False-positive alerts	Unnecessary operational alerts should be minimized	Precision
Pharmaceutical inspection	Missed defects/ false rejection	Both detection completeness and false-decision control matter	Recall/Precision
Microplastics detection	Missed small particles	Detection completeness and localization are important	Recall/mAP
CCPP monitoring	Train–test inconsistency	Candidate should preserve performance beyond training data	Generalization Stability

4.2. Criterion Roles

DAMSA distinguishes four evidence roles to prevent directly measured, derived, mandatory, and descriptive information from being treated as equivalent evidence in the decision process:

1. Core criteria (E_{core}) are directly measured performance criteria used in Pareto screening and SMS scoring. The vision scenarios use mAP, Precision, and Recall; CCPP monitoring uses Testing Performance and Generalization Stability.
2. Composite criteria (E_{comp}) are derived from core measurements and represent additional decision preferences. F1-score is treated as a composite balance criterion and is excluded from Pareto screening.
3. Hard constraints (H) represent non-negotiable measured requirements, such as minimum Recall or maximum latency, and are applied before ranking.
4. Descriptive measurements (E_{desc}) are retained for transparency but do not enter Pareto screening or SMS scoring. CCPP Training Performance is treated descriptively because it contributes to the Generalization Stability calculation.

These roles determine how each measurement participates in the framework. Core criteria enter both Pareto screening and reference scoring, composite criteria may enter reference scoring but are excluded from Pareto dominance when algebraically derived from core measurements, hard constraints determine candidate feasibility before ranking, and descriptive measurements are retained only for interpretation or calculation support. This separation reduces double-counting and makes candidate eligibility and ranking logic reproducible.

4.3. Criterion Normalization

Let x_{ij} denote the observed value of candidate M_i under active criterion e_j. The percentage-based benefit criteria used in this study have fixed theoretical bounds from 0 to 100 and are normalized as:

$$\hat{x}_{ij} = \frac{x_{ij}}{100}, j \in E_{active} \quad (4)$$

Thus, $\hat{x}_{ij} \in [0, 1]$, with larger values indicating greater deployment suitability. Future cost-type criteria must first be transformed into corresponding suitability measures before entering the SMS calculation. Because the active criteria used in the present experiments are

percentage-based benefit measures with common theoretical bounds of 0–100, fixed-bound normalization is used instead of sample-dependent min–max normalization. This preserves the absolute interpretation of the reported performance values and avoids changing a candidate's normalized value merely because another candidate is added to or removed from the comparison set.

4.4. Feasibility and Pareto Screening

When measured mandatory requirements are available, the feasible candidate set is:

$$\mathcal{F}(D) = \{M_i \in \mathcal{M} : M_i \text{ satisfies every } h \in H\} \quad (5)$$

A hard constraint activates only when the deployment scenario includes a measurable, non-negotiable requirement that can be evaluated consistently across all candidates, such as a minimum Recall threshold or a maximum latency limit. Candidates violating any active hard constraint are removed before preference-based ranking. When no such measured requirement is available, no candidate is excluded at the feasibility stage and $\mathcal{F}(D) = \mathcal{M}$. Pareto dominance is then evaluated using only the core criteria. Candidate M_a dominates M_β when:

$$M_a \succ M_\beta \Leftrightarrow \hat{x}_{aj} \geq \hat{x}_{\beta j} \forall\, j \in E_{core}, \quad (6)$$

and

$$\hat{x}_{aj} > \hat{x}_{\beta j} \text{for at least one } j \in E_{core} \quad (7)$$

The Pareto-efficient set is therefore:

$$\mathcal{P}(D) = \{M_i \in \mathcal{F}(D):\ \nexists\, M_k \in \mathcal{F}(D) \text{ such that } M_k \succ M_i\} \quad (8)$$

F1-score is excluded from the dominance test because it is derived from Precision and Recall. Formal DAMSA ranking is applied to candidates in $\boldsymbol{\mathcal{P}}(D)$. Pareto screening is applied only to the core criteria because these represent directly measured evidence. Its purpose is candidate control rather than final preference determination: a candidate is removed only when another feasible candidate is at least as good on every core criterion and strictly better on at least one. Non-dominated candidates are retained even when they involve different performance trade-offs, leaving those trade-offs to be resolved by the documented deployment priorities.

4.5. Deployment-Priority Profile

For deployment scenario D, the documented reference priority vector is defined over the active criteria:

$$w_{ref}(D) = (w_1, \ldots, w_n), j \in E_{active} \quad (9)$$

subject to:

$$w_j(D) \geq 0, \sum_{j \in E_{active}} w_j(D) = 1 \quad (10)$$

The reference vector ($w^{ref}(D)$) represents one documented deployment-priority profile rather than a statistically learned or universally optimal set of weights. Its role is to provide an interpretable reference operating point for the scenario. The relative weights are assigned from the stated operational concern and criterion priorities of the deployment case, while uncertainty in the exact preference values is addressed separately through the admissible policy region ($\mathcal{W}(D)$) and first-rank acceptability analysis. Consequently, the final interpretation does not rely exclusively on a single fixed weight vector.

4.6. Reference SMS Index

The reference SMS Index is calculated for every Pareto-efficient candidate using the active scoring criteria:

$$SMS_i(D,\ w) = 100 \sum_{j \in E_{active}} w_j(D)\hat{x}_{ij},\ M_i \in \mathcal{P}(D) \tag{11}$$

Under the present benefit-type formulation, ($SMS_i(D,w)$) is a convex weighted aggregation of normalized criterion values. Because $\left(\hat{x}_{ij} \in [0,1]\right)$, $\left(w_j \geq 0\ \right)$, and $\left(\sum_j w_j = 1\right)$, the resulting SMS Index lies between 0 and 100. A higher value indicates greater suitability under the specified reference priority profile. The index is compensatory: a weaker value on one active criterion may be offset by stronger values on other criteria according to their assigned weights. For this reason, feasibility and Pareto screening are performed before formal ranking rather than relying on weighted aggregation alone.

The reference recommendation is:

$$M^*_{ref} = arg \max_{M_i \in \mathcal{P}(D)} SMS_i\left(D,\ w^{ref}\right) \tag{12}$$

The factor of 100 expresses the SMS Index on a 0–100 scale. The methodological contribution lies in constructing, controlling, screening, and stress-testing the deployment decision surrounding this transparent reference aggregation mechanism. Accordingly, superiority of DAMSA-SMS is not defined as producing a different first-ranked candidate from conventional weighted ranking. Its added decision value is evaluated through the additional information provided about candidate eligibility, Pareto efficiency, composite-criterion dependence, preference sensitivity, and metric sensitivity.

4.7. F1-Score Ablation

For the four vision scenarios, the active set without F1-score is:

$$E_{abl} = E_{active} \setminus \{F1\} \tag{13}$$

The remaining weights are renormalized as:

$$w_j^{-F1} = \frac{w_j}{1 - w_{F1}}, j \in E_{abl} \tag{14}$$

The ablated SMS value is then:

$$SMS_i^{-F1} = 100 \sum_{j \in E_{abl}} w_j^{-F1}\hat{x}_{ij} \tag{15}$$

The reference winner and top-two margin are compared before and after F1 removal. An unchanged winner indicates that the composite criterion contributes balance information without determining the recommendation. This ablation is included because F1-score is algebraically derived from Precision and Recall and may therefore partially duplicate information already represented by the core criteria. The analysis tests whether the reference recommendation depends materially on this composite criterion. If the selected candidate remains unchanged after F1-score is removed and the remaining weights are renormalized, the recommendation is interpreted as not being driven by composite-metric inclusion.

4.8. Preference-Robustness Analysis

A single reference vector may not represent every defensible deployment policy. The admissible deployment-policy region is defined as:

$$\mathcal{W}(D) = \left\{ w \,:\, w \geq 0,\ \sum_j w_j = 1,\ A_D\, w \leq b_D \right\} \tag{16}$$

where $A_D w \leq b_D$ represents the scenario-specific priority relations. For example, a false-alert-sensitive defense policy may impose $w_{Precision} \geq w_{mAP}$, $w_{Precision} \geq w_{Recall}$, while the composite criterion may be limited by $w_{F1} \leq 0.25$.

The admissible region $\mathcal{W}(D)$ therefore represents a set of defensible deployment-priority profiles rather than unrestricted arbitrary weights. The imposed relations encode scenario-specific priority statements derived from the documented operational context. For example, the defense constraints above ensure that Precision is not assigned less importance than mAP or Recall when false-positive operational alerts constitute the primary deployment concern.

For the vision scenarios, B accepted Monte Carlo samples w^b are drawn from $\mathcal{W}(D)$. Estimated first-rank acceptability is:

$$\hat{a}_i^1 = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}[\, i = arg \max_{k \in \mathcal{P}(D)} SMS_k\left(D, w^b\right)] \tag{17}$$

First-rank acceptability represents the proportion of the defined admissible priority region in which a candidate ranks first. It measures preference sensitivity, not prediction confidence, deployment-success probability, or statistical superiority. For the two-criterion CCPP case, preference shares and the switching boundary are calculated analytically.

4.9. Final Recommendation and Recommendation Trace

For the vision scenarios, a recommendation is considered preference-stable when the same candidate ranks first across all accepted priority profiles. When multiple candidates become preferred under different admissible priorities, the reference winner and competitive alternatives are reported conditionally. For CCPP monitoring, the same interpretation is obtained analytically over the defined priority interval.

For interpretation, three recommendation states are distinguished. A recommendation is termed preference-stable when the same candidate ranks first throughout the admissible policy region. It is termed conditionally preferred when the reference winner remains dominant across most of the region but one or more alternatives become preferred under admissible priority profiles. A recommendation is interpreted as preference-sensitive when substantial portions of the admissible region support different candidates. Metric robustness is reported separately and is not inferred from preference stability.

The recommendation trace records the deployment context and risk-to-criterion mapping; criterion roles and constraints; feasible and Pareto-efficient candidates; reference priorities, SMS scores, and ranking; F1-ablation and preference-robustness results; metric-perturbation behavior; and the resulting preference-stable recommendation or conditional shortlist.

4.10. DAMSA Procedure

Algorithm 1 provides the executable logical sequence of DAMSA, while Table 4 expands each operation for reproducibility and links the procedural steps to the corresponding equations. The procedure does not retrain or modify any candidate model; all operations are performed on the post-training evaluation evidence and documented

deployment information.

Algorithm 1. Deployment-Aware Model Selection Algorithm (DAMSA).

Input: deployment scenario (D); trained candidate set ($\mathcal{M}$); candidate evaluation matrix (X); criterion metadata; reference priority profile (w^{ref}) or ordinal priority relations; optional hard constraints (H); number of accepted preference samples (B).
Output: feasible set ($\mathcal{F}(D)$); Pareto-efficient set ($\mathcal{P}(D)$); reference ranking; F1-ablation result where applicable; first-rank acceptability; metric-perturbation result; final recommendation trace.

Algorithm

1. Construct the deployment specification ($D = (O,R,C)$).
2. Apply ($\Phi(D)$) to define (E_{core},E_{comp},H,E_{desc}).
3. Construct ($E_{active} = E_{core} \cup E_{comp}$).
4. Remove criteria that are unavailable or lack documented deployment relevance.
5. Normalize retained active criteria.
6. Apply hard constraints to construct ($\mathcal{F}(D)$).
7. Apply Pareto screening on (E_{core}) to construct ($\mathcal{P}(D)$).
8. Define the reference priority profile (w^{ref}).
9. Compute the reference SMS Index for all (M_i *in* $\mathcal{P}(D)$).
10. **If** a composite F1-score is present, perform F1-score ablation and recompute the reference ranking.
11. Construct the admissible preference region ($\mathcal{W}(D)$).
12. **If** the preference space is sampled numerically, generate (B) accepted priority vectors and estimate first-rank acceptability; **otherwise**, calculate the preference boundary analytically.
13. Perturb directly measured criteria over the declared bounds, recompute derived criteria, Pareto membership, and the reference recommendation.
14. Classify the recommendation as stable, conditional, or sensitive using the preference and metric-robustness results.
15. **Return** the Pareto set, reference ranking, robustness results, and recommendation trace.

Table 4. Detailed operational steps corresponding to Algorithm 1.

Step	Procedure
1	Define $D = (O, R, C)$, including the operational objectives, risks, and measured constraints.
2	Translate D into the role-partitioned sets E_{core}, E_{comp}, H, and E_{desc} using Equation (2).
3	*Construct* $E_{active} = E_{core} \cup E_{comp}$ using Equation (3), while retaining E_{desc} only for descriptive reporting.
4	Remove active criteria that are unavailable for one or more candidates or lack documented deployment relevance.
5	Normalize the retained percentage-based active criteria using Equation (4).
6	Apply measured hard constraints and construct $\mathcal{F}(D)$ using Equation (5); if no hard constraint is active, set $\mathcal{F}(D) = \mathcal{M}$.
7	Apply Pareto screening using only E_{core} and construct $\mathcal{P}(D)$ using Equations (6)–(8).

Table 4. *Cont.*

Step	Procedure
8	Establish and document the reference deployment-priority profile over E_{active} using Equations (9) and (10).
9	Calculate reference SMS values and identify M^*_{ref} using Equations (11) and (12).
10	For vision scenarios, construct E_{abl}, renormalize weights using Equation (14), and calculate leave-F1-out scores using Equation (15).
11	Construct the admissible policy region $\mathcal{W}(D)$ using Equation (16) and sample B accepted priority vectors for the vision scenarios.
12	Recalculate the SMS ranking for every accepted vector and estimate first-rank acceptability using Equation (17).
13	Perturb the directly measured performance criteria over the declared bounded ranges, recalculate derived criteria, Pareto membership, and the reference recommendation, and record the smallest perturbation associated with a change.
14	For a low-dimensional analytical case, calculate the exact preference-switching boundary.
15	Return the Pareto set, reference ranking, ablation result, acceptability values, recommendation interpretation, and full recommendation trace.

4.11. Computational Requirement

Let m denote the number of candidate models, $n = |E_{active}|$ the number of active scoring criteria, $c = |E_{core}|$ the number of core criteria, and B the number of accepted priority profiles. Normalization and reference scoring require $O(mn)$, direct Pareto screening requires $O(m^2c)$, and Monte Carlo preference analysis requires $O(Bmn)$. These costs apply only to the post-training decision process and are small relative to model training for the candidate-set sizes considered in this study.

5. Experimental Design and Comparative Evaluation Protocol

5.1. Study Design

The experimental evaluation examined whether DAMSA-SMS could provide traceable and robust post-training model recommendations across different engineering applications. All candidate models were trained and benchmarked before DAMSA-SMS was applied; the framework therefore does not modify model architectures, training procedures, hyperparameters, or learned parameters.

Validation was conducted at three levels. First, decision consistency was examined through feasibility and Pareto screening followed by comparison between benchmark signals and the reference SMS recommendation. Second, preference robustness was evaluated through F1-score ablation and first-rank acceptability across admissible deployment-priority profiles. Third, performance robustness was evaluated through bounded perturbations of the measured benchmark values. Together, these analyses examine whether the recommendation is supported by the candidate evidence and whether it remains stable under changes in decision priorities and measured performance.

To strengthen comparative validity, the revised evaluation also distinguishes between the reference-ranking result and the additional decision information produced by DAMSA-SMS. Conventional decision procedures are evaluated using the same candidate evidence and, where applicable, the same reference-priority assumptions. The comparison therefore examines not only whether different procedures identify the same nominal winner, but

also whether agreement at the reference operating point is sufficient to establish robustness under alternative priorities and bounded benchmark variation.

5.2. Deployment Scenarios

Five engineering scenarios were evaluated: smart traffic monitoring, defense vehicle surveillance, pharmaceutical tablet inspection, microplastics detection, and Combined Cycle Power Plant (CCPP) monitoring. The scenarios were selected to provide heterogeneous decision conditions rather than five repetitions of the same ranking problem. The four vision cases represent different operational error priorities, including missed detections, false-positive alerts, competing defect-control consequences, and incomplete small-object detection, whereas the CCPP case introduces a regression setting in which held-out performance must be considered together with train–test consistency. The scenarios also contain different Pareto structures and preference conditions, ranging from a single Pareto-efficient candidate in microplastics detection to multiple efficient alternatives with meaningful trade-offs in traffic, defense, pharmaceutical inspection, and CCPP monitoring. Their purpose is therefore to examine procedural transferability across task types, risk structures, and decision conditions rather than to imply that a universal criterion set applies to all applications. For the four vision scenarios, the candidate set includes multiple YOLO generations, namely YOLOv5s, YOLOv6s, YOLOv7, YOLOv8s, YOLOv9s, YOLOv11s, and YOLOv12s, representing successive developments in real-time object detection architectures [17,25–27]. The deployment scenarios and corresponding candidate-model evidence are summarized in Table 5.

Table 5. Deployment scenarios and candidate-model evidence.

Scenario	ML Task	Dataset/Evidence	Candidate Models	Main Deployment Concern
Smart traffic monitoring	Multi-class vehicle detection	5000 annotated images; 8 vehicle classes	YOLOv5s, v6s, v7, v8s, v9s, v11s, v12s	Missed vehicles and downstream traffic reliability
Defense vehicle surveillance	Military-vehicle detection	MVRSD [28]; 3000 images; 32,626 targets across 40 scenarios	YOLOv5s, v6s, v7, v8s, v9s, v11s, v12s	False-positive operational alerts
Pharmaceutical tablet inspection	Tablet-defect detection	1500 images under different lighting conditions	YOLOv5s, v6s, v7, v8s, v9s, v11s, v12s	Missed defects and unnecessary rejection decisions
Microplastics detection	Small-object detection	400 experimental images	YOLOv5s, v6s, v7, v8s, v9s, v11s, v12s	Missed microplastic particles
CCPP monitoring	Regression	Eight years of power-plant operating data	AdaBoost DT, Random Forest Regression, Gradient Boosting, KNN Regression, KNN AdaBoost, Kernelized SVM, Linear Regression, Linear SVM, and Bagging DT	Testing performance and generalization stability

Within each deployment scenario, candidate models were compared using a common dataset partition and a common evaluation protocol so that the decision framework operated on directly comparable candidate evidence. DAMSA-SMS does not alter model architectures, learned parameters, training procedures, or benchmark outputs; it is applied only after the candidate evaluations have been completed. Consequently, any differences in the final deployment interpretation arise from the documented decision structure and robustness analysis rather than from retraining a candidate under different conditions for the selection stage. The four vision datasets were divided into 90% training, 5% validation, and 5% testing. Candidate YOLO variants were trained for 150 epochs in Google Colab

under common scenario-specific conditions, and mAP, Precision, Recall, and F1-score were retained as benchmark evidence.

The CCPP candidate set represents several regression families with different modeling characteristics. AdaBoost DT and Gradient Boosting are sequential ensemble approaches, whereas Random Forest Regression and Bagging DT combine multiple decision trees through parallel ensemble aggregation. KNN Regression uses neighborhood-based prediction, while KNN AdaBoost extends this approach through boosting. Kernelized SVM models nonlinear relationships through kernel functions, whereas Linear SVM and Linear Regression provide linear modeling alternatives. The candidate set therefore represents heterogeneous regression approaches rather than minor variants of a single model family [29,30].

The CCPP dataset used the same 90:5:5 split. Its available measurements were Training Performance, Testing Performance, and Generalization Stability. For CCPP monitoring, Generalization Stability was derived from the absolute difference between Training and Testing Performance:

$$Generalization\ Gap_i = |Training_i - Testing_i| \quad (18)$$

$$Generalization\ Stability_i = 100 - Generalization\ Gap_i \quad (19)$$

5.3. Decision Matrices and Criterion Roles

For the vision scenarios, mAP, Precision, and Recall are treated as the core criteria for Pareto screening and SMS scoring. F1-score is included in the full SMS calculation as a measure of balance between Precision and Recall, but it is excluded from Pareto screening because it is derived from these two metrics. For CCPP monitoring, Testing Performance and Generalization Stability are used as the core criteria. Training Performance is reported only as supporting information because it is already used to calculate Generalization Stability and is not treated as a separate deployment benefit. The criterion structure is therefore determined before candidate ranking and is held fixed when comparing alternative decision procedures within a scenario. This prevents baseline methods from being advantaged or disadvantaged through different criterion definitions.

Table 6 reports reference deployment-priority profiles rather than optimized or statistically estimated weights. These profiles provide a transparent reference operating point derived from the documented operational emphasis of each scenario. The weights should therefore be interpreted as representative decision assumptions rather than universal or uniquely correct values.

Table 6. Reference SMS priority profiles.

Scenario	mAP	Precision	Recall	F1-Score	Testing Performance	Generalization Stability	Reference Interpretation
Traffic monitoring	0.30	0.20	0.25	0.25	—	—	Balanced localization, detection completeness, and downstream reliability
Defense surveillance	0.25	0.40	0.15	0.20	—	—	Strong false-positive control
Pharmaceutical inspection	0.25	0.30	0.25	0.20	—	—	Balanced defect identification with slightly greater emphasis on false-decision control
Microplastics detection	0.30	0.20	0.30	0.20	—	—	Localization and detection completeness for small particles
CCPP monitoring	—	—	—	—	0.7778	0.2222	Testing-focused profile with explicit consideration of generalization stability

Table 7 extends the fixed reference profiles in Table 6 by defining broader ordinal priority relations for the robustness analysis. Thus, Table 6 provides one nominal reference operating point, whereas Table 7 defines a family of admissible deployment policies. For example, the defense reference profile assigns the largest weight to Precision because false-positive operational alerts constitute the primary deployment concern. The robustness analysis does not require the exact reference value of 0.40 to remain fixed; it requires only that Precision remain at least as important as mAP and Recall within the declared policy region. The same principle is applied to the remaining scenarios.

Table 7. Operational justification for reference priority profiles and admissible policy relations.

Scenario	Operational Concern	Reference Emphasis	Admissible Policy Relation
Traffic	Missed vehicles and downstream reliability	Recall	$w_{Recall} \geq w_{Precision}$
Defense	False-positive alert burden	Precision	$w_{Precision} \geq w_{mAP}$; $w_{Precision} \geq w_{Recall}$
Pharmaceutical	Missed defects and false rejection	Precision and Recall; F1 secondary	$w_{mAP}, w_{Precision}, w_{Recall} \geq w_{F1}$
Microplastics	Incomplete particle detection	Recall and mAP	$w_{Recall} \geq w_{Precision}$; $w_{mAP} \geq w_{Precision}$
CCPP	Testing performance versus stability	Testing Performance	$w_{Testing} \geq w_{GS}$

5.4. CCPP Decision Profiles

The CCPP case is reported through two complementary deployment interpretations. A testing-performance-focused profile selects the candidate with the highest held-out Testing Performance. A testing-generalization profile combines Testing Performance and Generalization Stability. After Training Performance is removed from the original 0.70-0.20-0.10 profile, the remaining weights are renormalized to $w_{Testing} = \frac{0.70}{0.90} = 0.7778$, $w_{Generalization} = \frac{0.20}{0.90} = 0.2222$.

These two profiles are reported deliberately because they represent different deployment questions. The testing-focused profile asks which candidate performs best on held-out data alone, whereas the testing-generalization profile asks which candidate provides the stronger compromise between held-out performance and train–test consistency. The comparison is therefore not intended to show that one profile is universally correct, but to make the policy dependence of the recommendation explicit.

5.5. DAMSA-SMS Experimental Procedure

The formal DAMSA operations defined in Table 4 were implemented experimentally through the ten higher-level stages summarized in Figure 2.

Steps 1–3 construct the deployment-specific decision model, Steps 4–6 control candidate eligibility and establish the reference ranking, Steps 7–9 evaluate composite-metric, preference, and performance sensitivity, and Step 10 integrates these results into the final recommendation. The experimental procedure therefore separates three questions: which candidates remain eligible for formal comparison, which candidate is preferred under the documented reference policy, and how strongly that recommendation depends on modeling assumptions or variation in the evidence. This separation is used consistently across all five scenarios.

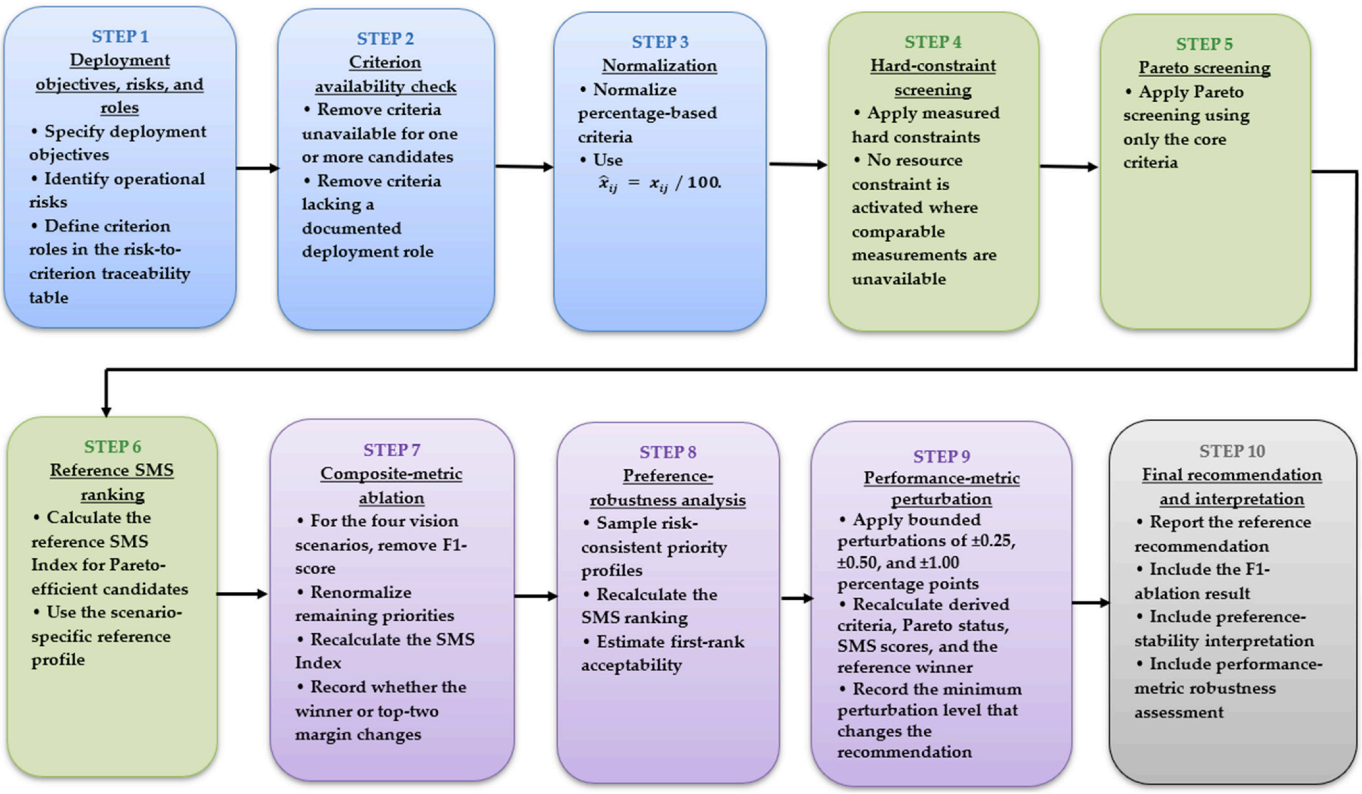


Figure 2. Flowchart of the DAMSA-SMS experimental procedure, showing deployment-context specification, criterion preparation, candidate screening, reference SMS ranking, composite-metric ablation, preference-robustness analysis, performance-metric perturbation, and final recommendation interpretation.

5.6. Comparative Baseline Protocol

To determine how the DAMSA-SMS reference recommendation compares with simpler and established decision procedures, a controlled comparative analysis was performed using benchmark-only selection, direct weighted-sum ranking (SAW), TOPSIS, Pareto-only screening, Pareto screening followed by weighted-sum ranking (Pareto + SAW), and the complete DAMSA-SMS procedure.

Benchmark-only selection identifies the candidate with the highest conventional primary benchmark for each scenario, namely mAP for the vision cases and Testing Performance for CCPP monitoring. SAW performs direct weighted aggregation of the normalized active criteria using the same reference-priority vector employed by the SMS Index, but without DAMSA's candidate-control and post-ranking robustness stages. TOPSIS provides an alternative MCDA baseline by ranking candidates according to their weighted relative closeness to the positive ideal and negative ideal solutions. Pareto-only analysis identifies the non-dominated candidate set using the same core criteria adopted in DAMSA but does not impose a preference-based ranking within that set. Pareto + SAW first restricts the candidate set to the non-dominated alternatives under the same core criteria used by DAMSA and subsequently applies the same weighted aggregation.

For fairness, all comparative procedures operate on the same scenario-specific candidate evidence, active criteria, criterion directions, and reference priorities wherever weights are required. SAW uses the fixed-bound normalization defined in Equation (4), whereas TOPSIS applies its conventional vector normalization before weighted ideal-solution analysis. No candidate model is retrained or assigned different benchmark evidence for any baseline. Consequently, the comparison evaluates differences in decision procedure rather than differences in predictive-model training or test data.

Agreement between DAMSA-SMS and a conventional ranking procedure is not treated as evidence that DAMSA-SMS is unnecessary, nor is disagreement treated automatically as evidence of superiority. Instead, the comparative analysis examines whether a single fixed ranking captures the same decision information as the full framework, particularly when preference variation and benchmark perturbation are considered.

5.7. Preference-Robustness Protocol

For the four vision scenarios, candidate priority vectors were generated from a uniform Dirichlet distribution, $w \sim Dirichlet(1,1,1,1)$, and retained only when they satisfied the scenario-specific relations in Table 7 together with the composite-weight restriction $w_{F1} \leq 0.25$.

Sampling continued until 500,000 accepted priority vectors were obtained for each vision scenario. A fixed random seed of 42 was used for reproducibility. For each accepted vector, the SMS Index was recalculated for every Pareto-efficient candidate, and the first-ranked candidate received one first-place count. First-rank acceptability was then estimated using Equation (17). The maximum Monte Carlo standard error of an estimated proportion with 500,000 accepted samples is approximately 0.071 percentage points; results are therefore reported to one decimal place. The large accepted-sample size was chosen to make Monte Carlo sampling error negligible relative to the reported one-decimal-place first-rank acceptability values. First-rank acceptability represents the share of the declared admissible preference region under which a candidate ranks first and should not be interpreted as prediction confidence or deployment-success probability.

For CCPP monitoring, preference shares and the switching point were calculated analytically because the two-criterion preference space can be represented by the Generalization Stability weight.

5.8. Performance-Metric Perturbation Protocol

Preference robustness varies the deployment priorities while holding benchmark evidence fixed. To separately examine sensitivity to variation in the measured performance values, directly measured criteria were perturbed by $\delta \in \{0.25, 0.50, 1.00\}$ percentage points. For each candidate M_i and directly measured criterion e_j, the perturbed value was defined as:

$$x_{ij}^{\pm\delta} = min\left(100, max\left(0, x_{ij} \pm \delta\right)\right) \quad (20)$$

At each perturbation magnitude, derived criteria, Pareto membership, SMS scores, and the reference recommendation were recalculated.

For the vision scenarios, F1-score was recalculated after Precision or Recall perturbation:

$$F1_i^{\delta} = 2\frac{P_i^{\delta} R_i^{\delta}}{P_i^{\delta} + R_i^{\delta}} \quad (21)$$

For CCPP monitoring, the original Training and Testing Performance values were perturbed, after which Generalization Stability was recalculated from the resulting train–test gap:

$$GS_i^{\delta} = 100 - \left| Training_i^{\delta} - Testing_i^{\delta} \right| \quad (22)$$

Two perturbation modes were examined: one-at-a-time perturbation, in which one measured criterion of one candidate was changed while other values remained fixed, and conservative coordinated stress, in which the reference winner was adversely perturbed while its strongest competitor was favorably perturbed within the same bound. The perturbation analysis is a deterministic decision-sensitivity test and should not be interpreted as a statistical confidence interval for the underlying model-performance estimates. Its purpose

is to determine how much bounded change in the reported evidence is sufficient to alter the deployment recommendation. Statistical uncertainty arising from repeated training or resampling is treated separately as a limitation of the present validation.

6. Results

6.1. Pareto Screening and Reference Recommendations

Pareto screening reduced the candidate set in all five scenarios. The efficient sets were YOLOv7, YOLOv8s, and YOLOv11s for traffic monitoring; YOLOv8s and YOLOv11s for defense surveillance; YOLOv7, YOLOv12s, and YOLOv9s for pharmaceutical inspection; YOLOv8s alone for microplastics detection; and AdaBoost DT and Random Forest Regression for CCPP monitoring.

Microplastics detection produced a single Pareto-efficient candidate because YOLOv8s led all three core criteria, as shown in Table 8. The remaining scenarios retained multiple efficient candidates representing genuine performance trade-offs requiring deployment-priority interpretation.

Table 8. Pareto-efficient candidates and reference recommendations.

Deployment Scenario	Pareto-Efficient Candidates	Candidates Screened Out	Reference Recommendation
Smart traffic monitoring	YOLOv7, YOLOv8s, YOLOv11s	4 of 7	YOLOv7
Defense vehicle surveillance	YOLOv8s, YOLOv11s	5 of 7	YOLOv8s
Pharmaceutical tablet inspection	YOLOv7, YOLOv12s, YOLOv9s	4 of 7	YOLOv7
Microplastics detection	YOLOv8s	6 of 7	YOLOv8s
CCPP monitoring	AdaBoost DT, Random Forest Regression	7 of 9	Random Forest under the Testing–Generalization reference profile

6.2. Smart Traffic Monitoring

The reference profile was [*mAP*, *Precision*, *Recall*, *F*1] = [0.30, 0.20, 0.25, 0.25].

Table 9 shows that, under the reference profile, YOLOv7 achieved the highest SMS Index (97.872), followed by YOLOv8s (97.400) and YOLOv11s (97.275) among Pareto-efficient candidates. Removing F1-score did not change the winner; the top-two margin changed only from 0.473 to 0.374 SMS points. YOLOv7 was therefore retained as the reference recommendation. From a practical traffic-monitoring perspective, the result indicates that YOLOv7 provides the strongest reference compromise for vehicle detection under the stated priority profile, but the recommendation should not be interpreted as universally invariant. The 92.8% first-rank acceptability shows that YOLOv7 remains preferred across most admissible deployment priorities, while the remaining preference share for YOLOv8s indicates that a different choice may become reasonable when greater emphasis is placed on detection completeness. The coordinated perturbation result further suggests that the deployment decision should be reviewed if benchmark differences between the leading candidates narrow or if updated evaluation evidence becomes available.

Table 9. Complete DAMSA-SMS results for smart traffic monitoring.

Model	mAP	Precision	Recall	F1	Pareto Status	Full SMS	DAMSA Rank	SMS Without F1	Ablation Rank	First-Rank Acceptability
YOLOv7	99.20	95.60	98.80	97.17	Efficient	97.872	1	98.107	1	92.8%
YOLOv9s	99.00	95.50	98.10	96.70	Dominated	97.500	-	97.767	-	-
YOLOv8s	99.10	94.10	99.00	96.40	Efficient	97.400	2	97.733	2	7.2%
YOLOv11s	98.80	96.80	96.50	96.60	Efficient	97.275	3	97.500	3	0.0%
YOLOv12s	98.50	95.20	96.80	95.90	Dominated	96.765	-	97.053	-	-
YOLOv5s	98.20	95.00	95.80	95.39	Dominated	96.258	-	96.547	-	-
YOLOv6s	83.50	81.10	68.90	74.50	Dominated	77.120	-	77.993	-	-

6.3. Defense Vehicle Surveillance

The reference profile was [mAP, Precision, Recall, F1] = [0.25, 0.40, 0.15, 0.20].

YOLOv8s achieved a reference SMS Index of 86.045 compared with 78.620 for YOLOv11s as shown in Table 10. Although YOLOv11s had the highest mAP (84.2%), YOLOv8s achieved 94.2% Precision, an 18.9-percentage-point advantage in the criterion emphasized for false-positive alert control, while accepting a 3.7-percentage-point reduction in mAP. F1 ablation did not change the recommendation. Operationally, this result demonstrates why reliance on mAP alone may be insufficient for defense surveillance. Although YOLOv11s provides the highest mAP, YOLOv8s offers substantially higher Precision, which is more closely aligned with the stated concern of reducing false-positive operational alerts. Its 100% first-rank acceptability and stability through the evaluated $\pm$1.00-percentage-point perturbations further indicate that the recommendation is not only deployment-driven but also comparatively robust within the declared decision conditions.

Table 10. Complete DAMSA-SMS results for defense vehicle surveillance.

Model	mAP	Precision	Recall	F1	Pareto Status	Full SMS	DAMSA Rank	SMS Without F1	Ablation Rank	First-Rank Acceptability
YOLOv8s	80.50	94.20	76.00	84.20	Efficient	86.045	1	86.506	1	100.0%
YOLOv11s	84.20	75.30	79.80	77.40	Efficient	78.620	2	78.925	2	0.0%
YOLOv5s	79.50	78.20	71.50	74.70	Dominated	76.820	-	77.350	-	-
YOLOv12s	77.10	80.00	68.30	73.60	Dominated	76.240	-	76.900	-	-
YOLOv7	78.30	77.00	68.80	72.60	Dominated	75.215	-	75.869	-	-
YOLOv9s	64.50	62.50	60.20	61.30	Dominated	62.415	-	62.694	-	-
YOLOv6s	48.20	45.00	39.50	42.07	Dominated	44.389	-	44.969	-	-

6.4. Pharmaceutical Tablet Inspection

The reference profile was [mAP, Precision, Recall, F1] = [0.25, 0.30, 0.25, 0.20].

YOLOv7 achieved the highest reference SMS Index (98.428), followed by YOLOv12s (97.895) and YOLOv9s (97.830), as shown in Table 11. The top-two reference margin was 0.533 points, and removing F1-score did not change the winner. For pharmaceutical inspection, the practical implication is that YOLOv7 provides the strongest reference balance under the stated defect-control priorities, but the decision retains a meaningful conditional alternative. YOLOv12s becomes preferred within part of the admissible policy region and under coordinated benchmark variation of $\pm$0.50 percentage points, indicating that the deployment choice may change when the relative importance of missed defects and false rejection decisions shifts or when the leading candidates become more similar in measured performance.

Table 11. Complete DAMSA-SMS results for pharmaceutical tablet inspection.

Model	mAP	Precision	Recall	F1	Pareto Status	Full SMS	DAMSA Rank	SMS Without F1	Ablation Rank	First-Rank Acceptability
YOLOv7	98.90	98.60	97.90	98.24	Efficient	98.428	1	98.475	1	91.2%
YOLOv12s	98.70	97.00	98.40	97.60	Efficient	97.895	2	97.969	2	8.6%
YOLOv9s	98.10	98.70	96.70	97.60	Efficient	97.830	3	97.888	3	0.1%
YOLOv5s	98.60	96.00	97.20	96.59	Dominated	97.069	-	97.188	-	-
YOLOv11s	98.80	95.80	97.30	96.50	Dominated	97.074	-	97.206	-	-
YOLOv8s	98.50	96.40	96.60	96.50	Dominated	96.995	-	97.119	-	-
YOLOv6s	91.10	77.50	75.60	76.50	Dominated	80.233	-	81.156	-	-

6.5. Microplastics Detection

The reference profile was [mAP, Precision, Recall, F1] = [0.30, 0.20, 0.30, 0.20].

In Table 12, YOLOv8s was the only Pareto-efficient candidate, having the highest mAP, Precision, and Recall. It therefore remained the reference recommendation with and without F1-score. In the microplastics scenario, the practical decision is more straightforward because YOLOv8s is the only Pareto-efficient candidate and leads all three core criteria. This reduces dependence on the exact preference profile and explains its 100% first-rank acceptability. However, the coordinated perturbation results show that the recommendation is not completely insensitive to benchmark variation, so continued validation remains important when model-performance differences are small or when new evaluation data become available.

Table 12. Complete DAMSA-SMS results for microplastics detection.

Model	mAP	Precision	Recall	F1	Pareto Status	Full SMS	DAMSA Rank	SMS Without F1	Ablation Rank	First-Rank Acceptability
YOLOv8s	81.40	80.50	81.50	80.99	Efficient	81.168	1	81.213	1	100.0%
YOLOv7	80.73	79.86	81.10	80.40	Dominated	80.601	-	80.651	-	-
YOLOv12s	73.40	69.90	77.80	73.60	Dominated	74.060	-	74.175	-	-
YOLOv9s	72.80	68.50	76.50	72.30	Dominated	72.950	-	73.113	-	-
YOLOv11s	77.20	66.50	66.70	66.59	Dominated	69.788	-	70.588	-	-
YOLOv5s	52.20	66.70	63.00	64.70	Dominated	60.840	-	59.875	-	-
YOLOv6s	22.90	20.20	20.00	20.09	Dominated	20.928	-	21.138	-	-

To provide a cross-scenario comparison, Figure 3 summarizes the reference SMS scores of the Pareto-efficient candidates retained across the five deployment scenarios. Figure 4 complements this comparison by showing the first-rank acceptability of the preferred and alternative candidates across the admissible deployment-priority regions.

6.6. Combined Cycle Power Plant Monitoring

Training Performance is retained in the table for descriptive completeness. The principal DAMSA-SMS decision uses Testing Performance and Generalization Stability. The testing-generalization reference profile is [0.7778, 0.2222].

In Table 13, AdaBoost DT achieved the highest Testing Performance (98.32%) and is therefore the testing-focused recommendation. Random Forest Regression achieved slightly lower Testing Performance (98.12%) but substantially higher Generalization Stability (99.68% versus 98.34%). Consequently, under the testing-generalization reference profile, Random Forest achieved the higher SMS value (98.467 versus 98.324).

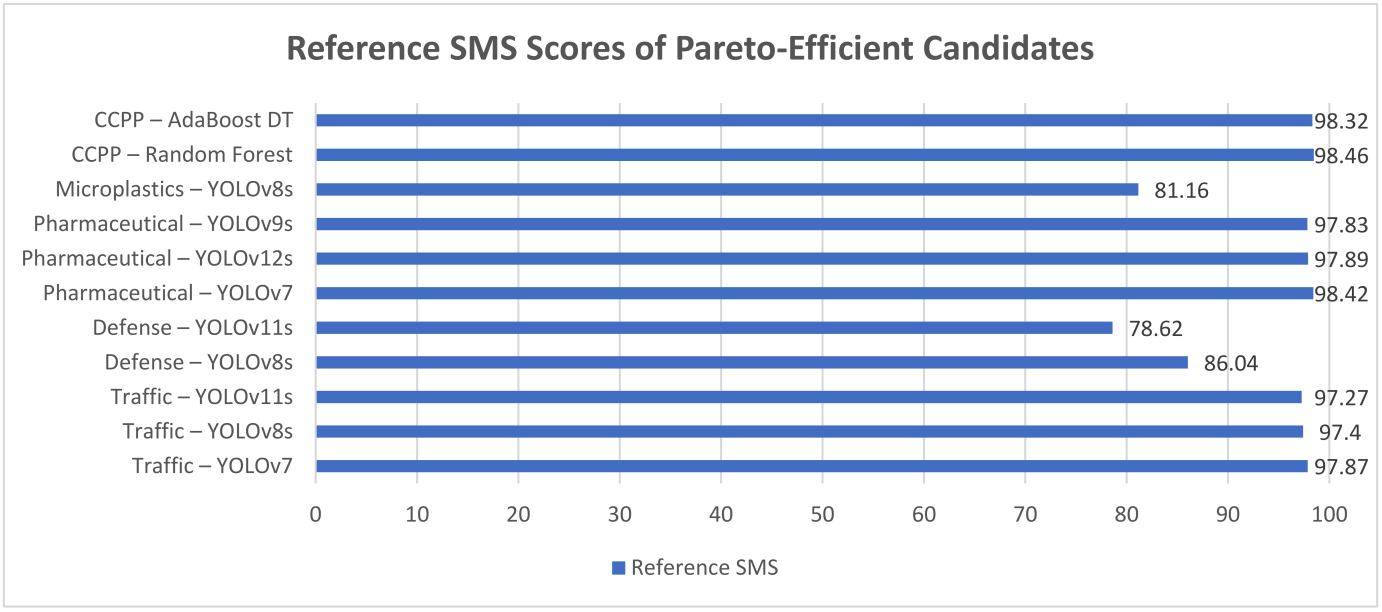


Figure 3. Reference SMS scores of Pareto-efficient candidates across the five deployment scenarios. Only candidates retained after Pareto screening are shown.

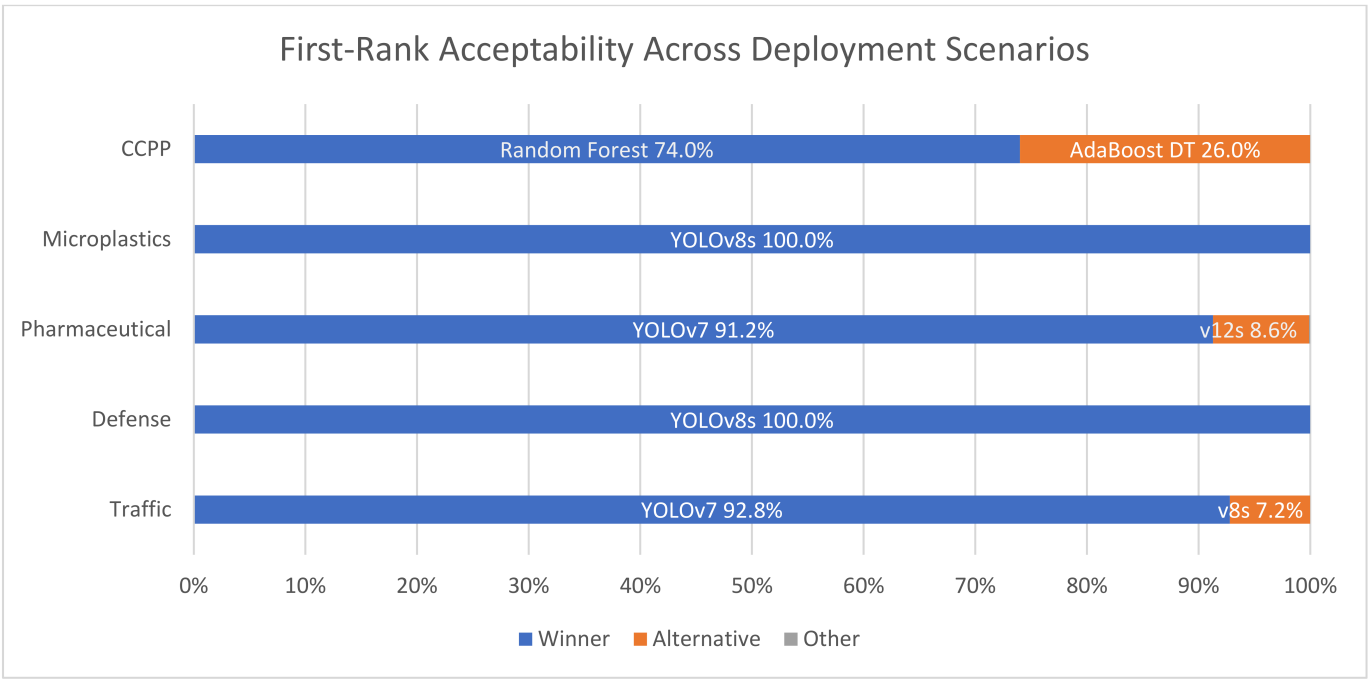


Figure 4. First-rank acceptability across the admissible deployment-priority regions. Vision results were estimated using 500,000 accepted priority vectors per scenario, whereas CCPP preference shares were calculated analytically. First-rank acceptability represents preference sensitivity rather than deployment-success probability.

Table 13. Complete CCPP deployment-decision results.

Model	Training	Testing	Generalization Stability	Pareto Status	Testing Rank	Testing-Generalization SMS	DAMSA-SMS Rank	First-Rank Acceptability
AdaBoost DT	99.98	98.32	98.34	Efficient	1	98.324	2	26.0%
Random Forest Regression	98.44	98.12	99.68	Efficient	3	98.467	1	74.0%
Gradient Boosting	99.92	98.24	98.32	Dominated	2	98.258	-	-
KNN Regression	99.99	97.41	97.42	Dominated	4	97.412	-	-
KNN AdaBoost	98.38	96.88	98.50	Dominated	5	97.240	-	-
Kernelized SVM	92.86	94.92	97.94	Dominated	6	95.591	-	-
Linear Regression	91.82	94.38	97.44	Dominated	7	95.060	-	-
Linear SVM	91.08	94.38	96.70	Dominated	7	94.896	-	-
Bagging DT	66.35	71.48	94.87	Dominated	9	76.678	-	-

Let g denote the Generalization Stability weight. The candidate-specific score functions are:

$$SMS_A(g) = (1 - g)T_A + gG_A, \tag{23}$$

$$SMS_R(g) = (1 - g)T_R + gG_R \tag{24}$$

The switching point is obtained by setting the two scores equal:

$$(1 - g)T_A + gG_A = (1 - g)T_R + gG_R \tag{25}$$

which gives:

$$g^* = \frac{(T_A - T_R)}{[(T_A - T_R) + (G_R - G_A)]} \tag{26}$$

Using the observed values for AdaBoost DT and Random Forest Regression in Equation (26), the switching point is ($g^* = 0.12987$), corresponding to a Generalization Stability weight of approximately 12.99%. Accordingly, AdaBoost DT is preferred when ($g < 12.99\%$), whereas Random Forest Regression is preferred when ($g > 12.99\%$). Across the admissible region ($0 \leq g \leq 0.50$), Random Forest Regression is therefore preferred over approximately 74.0% of the region and AdaBoost DT over approximately 26.0%. The resulting analytical preference-switching boundary is illustrated in Figure 5.

For CCPP monitoring, the practical value lies in converting a model-ranking problem into an explicit policy-dependent deployment rule. AdaBoost DT is preferable when held-out Testing Performance is given nearly exclusive priority, whereas Random Forest becomes preferable once the importance assigned to Generalization Stability exceeds 12.99%. This threshold gives operators a transparent basis for choosing between the two candidates according to how strongly train-test consistency should influence deployment. The local sensitivity to ± 0.25-percentage-point benchmark variation also indicates that the decision should be revisited when updated evaluation evidence becomes available.

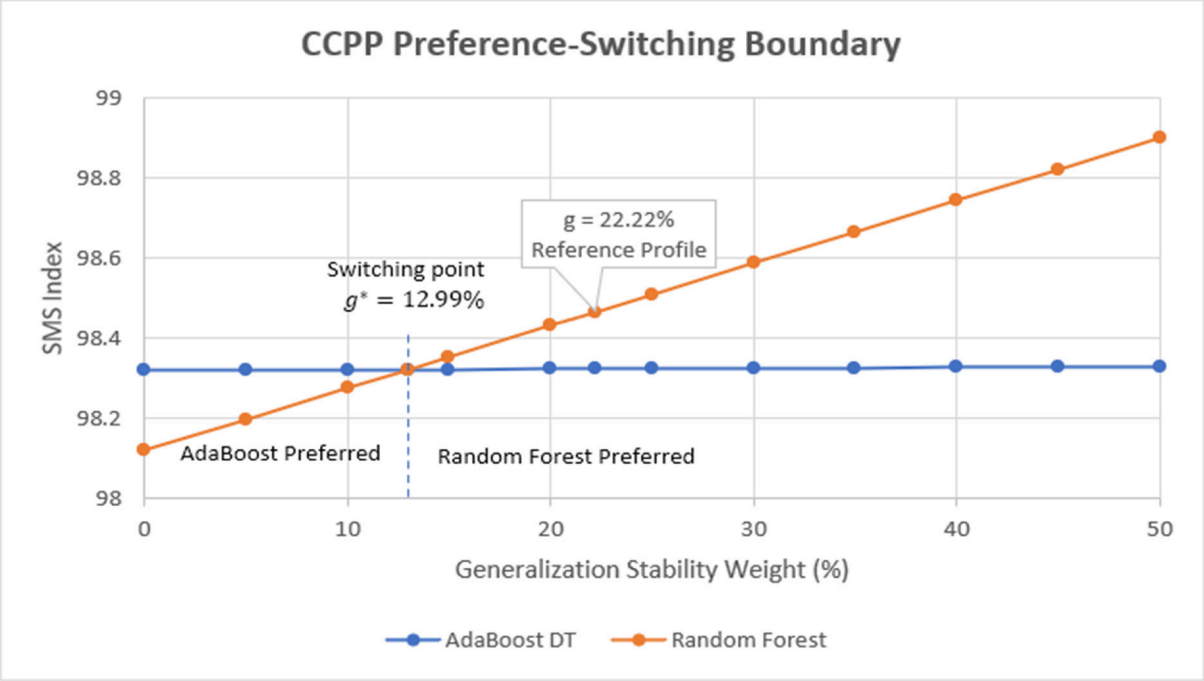


Figure 5. Analytical preference-switching boundary for CCPP monitoring. AdaBoost DT is preferred when the Generalization Stability weight is below 12.99%, whereas Random Forest Regression is preferred above this threshold within the admissible region. The reference profile assigns 22.22% weight to Generalization Stability.

6.7. Performance-Metric Perturbation Analysis

Bounded performance-metric perturbation was used to evaluate the stability of the reference recommendations under one-at-a-time and coordinated adverse changes in the benchmark evidence.

Table 14 reveals scenario-dependent differences in metric robustness, ranging from strong stability in defense surveillance to greater local sensitivity in traffic and CCPP monitoring. These findings complement first-rank acceptability by distinguishing sensitivity to benchmark variation from sensitivity to deployment priorities.

Table 14. Robustness of DAMSA-SMS recommendations under bounded performance-metric perturbations.

Scenario	Reference Winner	One-at-a-Time ±0.25	One-at-a-Time ±0.50	One-at-a-Time ±1.00	Conservative Stress Result	Interpretation
Traffic monitoring	YOLOv7	Stable	Stable	Switch possible to YOLOv9s	Switches to YOLOv9s at ±0.25	Sensitive to coordinated adverse variation
Defense surveillance	YOLOv8s	Stable	Stable	Stable	Stable through ±1.00	Strongly metric-robust
Pharmaceutical inspection	YOLOv7	Stable	Stable	Stable	Switches to YOLOv12s at ±0.50	Moderately metric-robust
Microplastics detection	YOLOv8s	Stable	Stable	Stable	Switches to YOLOv7 at ±0.50	Moderately metric-robust
CCPP monitoring	Random Forest	Switch possible	Switch possible	Switch possible	Switches to AdaBoost at ±0.25	Locally sensitive

6.8. Comparative Decision-Procedure Analysis

The reference recommendations were compared with benchmark-only selection, direct SAW ranking, TOPSIS, Pareto-only and Pareto + SAW under the controlled conditions described in Section 5.6. The purpose of this comparison was not to require DAMSA-SMS to produce a different nominal winner, but to determine whether conventional reference-ranking agreement provides the same information about deployment robustness as the complete framework.

Table 15 shows that the comparative value of DAMSA-SMS cannot be assessed solely by counting how often its nominal reference winner differs from a conventional ranking procedure. Two distinct behaviors were observed. In defense surveillance and CCPP monitoring, deployment-oriented multi-criteria evaluation changes the recommendation relative to selection based on the conventional primary benchmark. In traffic monitoring, pharmaceutical inspection, and microplastics detection, the nominal reference recommendation is consistent with the conventional ranking procedures, but the post-ranking analyses reveal substantially different levels of recommendation robustness.

Table 15. Comparative decision outcomes and robustness information across deployment scenarios.

Scenario	Conventional Decision Comparison	DAMSA-SMS Reference Recommendation	Preference Robustness	Metric Robustness	Deployment Interpretation
Traffic monitoring	Benchmark, SAW, TOPSIS and Pareto + SAW support YOLOv7; Pareto-only retains YOLOv7, YOLOv8s and YOLOv11s	YOLOv7	YOLOv7 92.8%; YOLOv8s 7.2%	Coordinated switch at ±0.25 pp	Nominal ranking is consistent, but the decision is not invariant to policy or benchmark variation
Defense surveillance	Benchmark-only favors YOLOv11s; SAW, TOPSIS and Pareto + SAW favor YOLOv8s; Pareto-only retains YOLOv8s and YOLOv11s	YOLOv8s	100%	Stable through ±1.00 pp	Operational Precision priority changes the benchmark-only decision and yields a strongly robust recommendation
Pharmaceutical inspection	Benchmark, SAW, TOPSIS and Pareto + SAW support YOLOv7; Pareto-only retains YOLOv7, YOLOv12s and YOLOv9s	YOLOv7	YOLOv7 91.2%; YOLOv12s 8.6%; YOLOv9s 0.1%	Coordinated switch at ±0.50 pp	Same nominal winner, but DAMSA identifies admissible conditional alternatives
Microplastics detection	Conventional procedures support YOLOv8s; Pareto-only leaves YOLOv8s as the sole efficient candidate	YOLOv8s	100%	Coordinated switch at ±0.50 pp	Preference-stable because the reference evidence is Pareto-unambiguous, but not fully invariant to metric variation
CCPP monitoring	Testing-only favors AdaBoost DT; multi-criteria reference procedures favor Random Forest; Pareto-only retains AdaBoost DT and Random Forest	Random Forest	RF ≈ 74%; AdaBoost ≈26%; switch at (g = 12.99%)	Switch possible at ±0.25 pp	The recommendation is explicitly policy-dependent and locally metric-sensitive

Pareto-only analysis also demonstrates why candidate screening and final recommendation are different decision stages. In four of the five scenarios, Pareto screening retains

multiple efficient candidates and therefore does not by itself identify a unique deployment choice. For example, three candidates remain efficient in both traffic monitoring and pharmaceutical inspection, two remain efficient in defense surveillance and CCPP monitoring, and only microplastics detection produces a single Pareto-efficient candidate. Pareto + SAW resolves these sets at the reference priority profile, but it still provides only a fixed reference ranking and does not characterize how the recommendation changes across alternative admissible priorities or bounded variation in benchmark evidence.

This distinction is especially evident in traffic monitoring and pharmaceutical inspection. Although the conventional reference procedures and DAMSA-SMS identify the same nominal winner, YOLOv8s becomes first-ranked across 7.2% of the admissible traffic-policy region, while YOLOv12s becomes first-ranked across 8.6% of the pharmaceutical-policy region. Moreover, the traffic recommendation can reverse under coordinated variation of only ± 0.25 percentage points and the pharmaceutical recommendation under ± 0.50 percentage points. Thus, agreement at one reference operating point does not establish that the deployment decision is invariant to defensible policy or measurement changes.

Defense surveillance illustrates a different form of added decision value. Benchmark-only selection favors YOLOv11s because it has the highest mAP, whereas the multi-criteria reference procedures select YOLOv8s under the false-alert-sensitive deployment profile. DAMSA-SMS additionally shows that this recommendation is preference-stable throughout the complete declared admissible policy region and remains unchanged under the evaluated perturbations through ± 1.00 percentage point. The framework therefore not only changes the benchmark-only recommendation but also establishes that the resulting deployment choice is comparatively robust to both policy and metric variation.

CCPP monitoring further demonstrates why a fixed ranking alone can be insufficient. Testing Performance alone favors AdaBoost DT, whereas the testing–generalization reference profile favors Random Forest. DAMSA-SMS identifies the exact preference-switching boundary at a Generalization Stability weight of 12.99%, below which AdaBoost DT is preferred and above which Random Forest is preferred. The decision is also locally sensitive to benchmark variation, with a switch possible at ± 0.25 percentage points. This result converts a simple model ranking into a conditional deployment rule that can be interpreted directly in terms of the operator's priority assigned to generalization stability.

The comparative analysis therefore does not indicate that DAMSA-SMS must produce a different reference winner from established ranking procedures in order to provide value. Instead, the results distinguish three questions that are otherwise easily conflated: which candidates are non-dominated, which candidate is preferred at a documented reference operating point, and whether that preference remains defensible when deployment priorities or benchmark evidence vary. DAMSA-SMS integrates these questions within a single recommendation trace.

7. Discussion

7.1. Overall Findings and Deployment Interpretations

The results demonstrate that post-training model selection can produce different forms of deployment recommendation depending on the relationship between benchmark performance and operational priorities. Across the five scenarios, DAMSA-SMS changed, resolved, confirmed, or conditioned the recommendation rather than systematically favoring a different model from benchmark-only evaluation.

Defense surveillance provides the clearest example of a deployment-driven change. Although YOLOv11s achieved the highest mAP, YOLOv8s was preferred because its substantially higher Precision better addressed the false-positive alert burden defined for the deployment scenario. Traffic monitoring and pharmaceutical inspection instead involved

competing benchmark leaders, for which DAMSA-SMS provided reference recommendations while retaining conditional alternatives under different admissible priorities. In microplastics detection, the framework confirmed the benchmark leader because one candidate dominated the core criteria. CCPP monitoring produced a conditional result, with the preferred model changing according to the relative priority assigned to Testing Performance and Generalization Stability.

The comparative baseline analysis further shows that agreement at a fixed reference operating point should not be interpreted as evidence that the deployment decision is robust. In several scenarios, conventional multi-criteria procedures and DAMSA-SMS identify the same nominal reference candidate; however, the post-ranking analyses reveal that the recommendation may still change under admissible priority variation or bounded changes in benchmark evidence. DAMSA-SMS therefore addresses a broader decision question than reference ranking alone: not only which candidate ranks first under one documented priority profile, but also how defensible that recommendation remains when the policy assumptions or measured evidence vary.

These outcomes show that the purpose of DAMSA-SMS is not to replace benchmark evaluation or necessarily select a different candidate. Benchmark results remain the empirical evidence for the decision; the framework provides a traceable process linking that evidence to deployment objectives, risks, criterion roles, and priorities.

7.2. Role of Pareto Screening and Composite-Metric Control

Pareto screening reduced the candidate space in every scenario without requiring criterion weights. This step removes candidates that are uniformly inferior under the core criteria while retaining alternatives that represent genuine performance trade-offs. The resulting decision therefore focuses preference-based ranking on candidates for which operational priorities can meaningfully affect the selection. The distinction between core and composite criteria also prevents derived indicators from being treated as independent evidence. F1-score was retained as a composite representation of Precision–Recall balance but excluded from Pareto screening because it is algebraically derived from those measurements. Removing F1-score did not change the reference recommendation in any vision scenario, indicating that its inclusion contributed balance information without determining the selected model.

7.3. Preference and Metric Robustness

The two robustness analyses address different sources of decision sensitivity. First-rank acceptability evaluates whether the recommendation depends on a particular deployment-priority profile while holding benchmark measurements fixed. Metric perturbation instead examines whether bounded variation in the reported benchmark evidence can alter the recommendation.

The results show that preference stability and metric robustness are not necessarily the same. Defense surveillance remained stable under both analyses, while traffic showed a strong preference for the reference model across different deployment priorities but was more sensitive to coordinated changes in the benchmark values. Pharmaceutical and microplastics recommendations were stable under isolated metric changes but could switch under coordinated adverse changes. CCPP was sensitive to both deployment priorities and small changes in benchmark performance, so its recommendation is better interpreted as conditional.

These two analyses therefore provide different views of recommendation robustness. Preference analysis examines policy sensitivity, while metric perturbation examines mea-

surement sensitivity. A model can be stable across different deployment priorities but still be sensitive to small changes in its measured performance, or vice versa.

7.4. Relationship with Existing MCDM Approaches

DAMSA-SMS is intended to complement rather than replace established multi-criteria decision-making methods. Methods such as AHP, SAW, TOPSIS, VIKOR, PROMETHEE, and ELECTRE provide established mechanisms for eliciting priorities or ranking alternatives once a decision structure has been specified [13–15].

The comparative experiments in this study support this complementary interpretation: established ranking procedures can identify the same reference candidate under a fixed decision profile, while DAMSA-SMS adds candidate-control and post-ranking evidence about whether that choice remains stable across admissible policies and bounded benchmark variation.

The contribution of DAMSA-SMS lies primarily in the surrounding deployment-decision architecture. Operational objectives and risks are translated into measurable criteria before ranking; direct, derived, descriptive, and mandatory evidence are assigned distinct roles; feasibility and Pareto screening control candidate eligibility; and the resulting recommendation is examined for sensitivity to both deployment priorities and benchmark variation. The SMS Index is therefore used as a transparent reference-ranking mechanism rather than proposed as a mathematically superior alternative to established MCDM operators.

7.5. Practical Implications

DAMSA-SMS operates after model training and benchmark evaluation, allowing the same trained candidates to be reconsidered when deployment priorities change without modifying their architectures or learned parameters. The recommendation trace makes the relationship between operational risks, evaluation criteria, candidate trade-offs, and the final selection explicit, which can support engineering review and deployment justification.

The framework can also accommodate deployment factors such as latency, memory consumption, model size, computational demand, and energy use when reliable and comparable measurements are available across candidates. Such measurements may be incorporated as active criteria or hard constraints depending on their operational role. Because these resource measurements were not consistently available in the present study, no experimental claim is made regarding improvements in computational efficiency, energy consumption, or emissions.

7.6. Limitations and Future Work

The underlying benchmark results are reported primarily as point estimates because repeated-run and resampling-based uncertainty estimates were not available consistently across all candidate models and scenarios. Bounded metric perturbation therefore evaluates the sensitivity of the recommendation to controlled changes in the reported benchmark values and should not be interpreted as a statistical confidence interval or as uncertainty arising from repeated model training or benchmark resampling.

First-rank acceptability shows how model preference changes across the defined range of deployment priorities. It should therefore be viewed as a measure of preference sensitivity rather than prediction confidence or the probability of successful deployment. It also does not imply that all possible deployment priorities are equally likely in practice. In addition, comparable deployment-efficiency measurements were not available for all candidate models, so these factors were not included as active decision criteria. Moreover, the comparative results are conditioned on the selected baseline implementations, normalization procedures, and reference-priority assumptions; alternative MCDA formulations

may produce different nominal rankings and should therefore be examined in broader comparative studies.

Future work will include repeated-run evaluation and bootstrap- or resampling-based uncertainty estimation to quantify variability in the underlying model-performance measurements more directly. It will also investigate more structured approaches for deployment-priority elicitation, including stakeholder- or expert-informed preference models, and incorporate comparable deployment-resource measurements such as latency, memory use, model size, computational demand, and energy consumption when these are consistently available. Finally, the selected models will be integrated into complete intelligent systems and evaluated under their intended deployment conditions to examine whether the decision recommendations remain valid in operational use.

8. Conclusions

This study proposed the System-Aware Model Selection (SMS) Framework, implemented through the Deployment-Aware Model Selection Algorithm (DAMSA), as a post-training approach for selecting trained machine-learning models according to deployment-specific objectives, risks, and priorities. The three-layer architecture integrates pre-ranking decision construction, candidate control and reference ranking, and post-ranking robustness analysis.

Evaluation across five engineering scenarios showed that DAMSA-SMS can support different forms of deployment decisions: resolving conflicting benchmark signals, changing a benchmark-based recommendation when operational priorities justify the trade-off, confirming an unambiguous benchmark leader, and identifying conditional recommendations when model preference depends on deployment policy. The comparative analysis further showed that agreement among conventional reference-ranking procedures does not necessarily imply recommendation robustness, because the preferred candidate may still change under admissible priority variation or bounded changes in benchmark evidence. Pareto screening, F1-score ablation, preference-robustness analysis, and bounded metric perturbation provide complementary evidence for interpreting the resulting recommendation rather than relying on a single reference ranking.

The contribution of DAMSA-SMS therefore lies in making the path from operational requirements to model recommendation explicit, traceable, and open to robustness analysis, rather than in the weighted reference score alone. Its added value is therefore not defined by producing a different nominal winner from established ranking procedures, but by integrating candidate control, preference sensitivity, and metric sensitivity into the deployment recommendation itself. Future work will extend the framework through repeated-run and uncertainty-aware validation, structured priority elicitation, comparable deployment-resource measurements, and evaluation of the selected models within complete intelligent systems.

Author Contributions: Conceptualization, V.S. and P.K.; methodology, V.S. and P.K.; software, V.S.; validation, V.S. and P.K.; formal analysis, V.S.; investigation, V.S.; data curation, V.S.; writing—original draft preparation, V.S.; writing—review and editing, V.S. and P.K.; visualization, V.S.; supervision, P.K.; project administration, P.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets used in this study are cited in the manuscript. Data that are not publicly available will be made available by the authors upon request.

Acknowledgments: The authors would like to thank the Department of Electrical and Computer Engineering, Faculty of Engineering, Thammasat School of Engineering, Thammasat University, for its valuable support and the resources provided for this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [CrossRef] [PubMed]
2. Dong, S.; Wang, P.; Abbas, K. A survey on deep learning and its applications. *Comput. Sci. Rev.* **2021**, *40*, 100379. [CrossRef]
3. Russell, S.J.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 4th ed.; Pearson: Hoboken, NJ, USA, 2021.
4. Kaur, R.; Singh, S. A comprehensive review of object detection with deep learning. *Digit. Signal Process.* **2023**, *132*, 103812. [CrossRef]
5. Hossin, M.; Sulaiman, M.N. Classification assessment methods. *Appl. Comput. Inform.* **2021**, *17*, 168–192. [CrossRef]
6. Hodson, T.O. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geosci. Model Dev.* **2022**, *15*, 5481–5487. [CrossRef]
7. Hendrycks, D.; Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In Proceedings of the 7th International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA, 6–9 May 2019.
8. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* 2019), Atlanta, GA, USA, 29–31 January 2019*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 220–229.
9. *NIST AI 100-1*; Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023.
10. Strubell, E.; Ganesh, A.; McCallum, A. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3645–3650.
11. Schwartz, R.; Dodge, J.; Smith, N.A.; Etzioni, O. Green AI. *Commun. ACM* **2020**, *63*, 54–63. [CrossRef]
12. Patterson, D.; Gonzalez, J.; Le, Q.V.; Liang, C.; Munguia, L.-M.; Rothchild, D.; So, D.R.; Texier, M.; Dean, J. Carbon emissions and large neural network training. *arXiv* **2021**, arXiv:2104.10350.
13. Madzík, P.; Falát, L. State-of-the-art on analytic hierarchy process in the last 40 years: Literature review based on Latent Dirichlet Allocation topic modelling. *PLoS ONE* **2022**, *17*, e0268777. [CrossRef] [PubMed]
14. Azhar, N.A.; Radzi, N.A.M.; Wan Ahmad, W.S.H.M. Multi-criteria Decision Making: A Systematic Review. *Recent Adv. Electr. Electron. Eng.* **2021**, *14*, e291021197468. [CrossRef]
15. Pandey, V.; Komal; Dincer, H. A review on TOPSIS method and its extensions for different applications with recent development. *Soft Comput.* **2023**, *27*, 18011–18039. [CrossRef]
16. Shafiq, M.; Gu, Z. Deep Residual Learning for Image Recognition: A Survey. *Appl. Sci.* **2022**, *12*, 8972. [CrossRef]
17. Terven, J.; Córdova-Esparza, D.; Romero-González, J.A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [CrossRef]
18. Tan, M.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 6105–6114.
19. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), Virtual Event, 3–7 May 2021.
20. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 10012–10022.
21. Verdecchia, R.; Sallou, J.; Cruz, L. A systematic review of Green AI. *WIREs Data Min. Knowl. Discov.* **2023**, *13*, e1507. [CrossRef]
22. Petchrompo, S.; Coit, D.W.; Brintrup, A.; Wannakrairot, A.; Parlikad, A.K. A review of Pareto pruning methods for multi-objective optimization. *Comput. Ind. Eng.* **2022**, *167*, 108022. [CrossRef]
23. Ezell, B.; Lynch, C.J.; Hester, P.T. Methods for Weighting Decisions to Assist Modelers and Decision Analysts: A Review of Ratio Assignment and Approximate Techniques. *Appl. Sci.* **2021**, *11*, 10397. [CrossRef]

24. Pelissari, R.; Oliveira, M.C.; Ben Amor, S.; Kandakoglu, A.; Helleno, A.L. SMAA methods and their applications: A literature review and future research directions. *Ann. Oper. Res.* **2020**, *293*, 433–493. [CrossRef]
25. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.
26. Jocher, G.; Qiu, J. *Ultralytics YOLO11*, Version 11.0.0; Ultralytics: Frederick, MD, USA, 2024.
27. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524. [CrossRef]
28. Bai, D.; Yu, Y.; Song, L.; Cheng, B.; Gao, H. Construction and validation of remote sensing image dataset for fine-grained detection of military vehicles. *J. Image Graph.* **2024**, *29*, 3564–3577. (In Chinese) [CrossRef]
29. Mienye, I.D.; Sun, Y. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access* **2022**, *10*, 99129–99149. [CrossRef]
30. Tufail, S.; Riggs, H.; Tariq, M.; Sarwat, A.I. Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms. *Electronics* **2023**, *12*, 1789. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.