Praca oryginalna *Original paper*

Calibrated two-threshold rejection option classification of shape-index extremes in quail eggs using weight and shell color

BETÜL DAĞOĞLU HARK[1], SEMA ALAŞAHAN[2], CAFER TAYYAR ATEŞ[2], MAHMUT ŞAMİL ŞAMLI[2]

[1]Department of Biostatistics, Faculty of Medicine, Fırat University, Elazığ, Türkiye
[2]Department of Animal Science, Faculty of Veterinary Medicine, University of Hatay Mustafa Kemal, Hatay, Türkiye

Received 04.02.2026 *Accepted 01.04.2026*

Dağoğlu Hark B., Alaşahan S., Ateş C. T., Şamli M. Ş.
Calibrated two-threshold rejection option classification of shape-index extremes in quail eggs using weight and shell color

Summary

The shape index (SI) is an indicator that quantifies egg shape and is associated with quality/hatching outcomes. However, direct measurement of SI in the field requires time and labor. The aim of this study is to develop and validate a probability-calibrated, two-threshold rejection option policy that can distinguish SI outliers (LOW/HIGH) in quail eggs using only weight and shell color, without measurement. In a dataset of 1,031 eggs (color: brown 61.9%, grayish white 38.1%), SI extremes were defined empirically ($Q_{0.25} \approx 67.20$, $Q_{0.72} \approx 76.20$). A classifier based on Hist Gradient Boosting was built using weight + color inputs. The probabilities obtained in the out-of-fold (OOF) scheme were corrected by sigmoid calibration. Two probability thresholds p_{lo}, p_{hi} were selected on the OOF distribution through a coverage-targeted search, and the final model was refined in two phases by "snapping" to the model's discrete probability support points. The final policy produced approximately 61.4% coverage within the extreme tails with $p_{lo} \approx 0.527$ and $p_{hi} \approx 0.761$, while the accepted subset provided balanced accuracy (BA) ≈ 0.72 (95% CI [0.67 – 0.77]). The proportion of outliers in the entire data was 53.1%, and the expected overall automation (acceptance × outlier rate) was calculated to be ≈ 32.6%. The reject region made the risk-coverage balance controllable by directing uncertain samples to measurement. Calibrated probabilities, OOF threshold setting, and a two-threshold rejection policy can reliably identify SI outliers based solely on two observable phenotypes. This approach provides a scalable pre-screening layer that reduces measurement burden in incubation and quality control lines. It can be further enhanced in the future with microstructure and line/age-stratified designs.

Keywords: shape index, quail, shell color, rejection option, probability calibration, out-of-fold, two-threshold policy

Quail eggs are brown, grayish white, or more rarely chalky or greenish in color, with speckled patterns of varying density and scale on top (1, 5, 22). The ground color and speckle color are not merely visual phenotypes: according to the literature, these indicators can significantly influence the shell ratio and internal quality measures (white index, yolk index, Haugh unit), while shell weight and thickness are decisive in maintaining egg integrity and internal quality (1). The shell is a functional barrier formed in the uterine section of the oviduct, which allows gas exchange with its porous microstructure while protecting the embryo from external influences. Although the direct effect of spot color on quality in table eggs may be limited, the effect of shell color on hatchability and chick hatch weight has been reported in multiple studies (5). Recommendations for field applications indicate that brown ground–brown speckle or grayish white ground–small black dot patterns may be preferred for incubation (15).

Despite growing evidence for the relationship between the shell ground color and the spot pattern with quality and hatchery output, a field-based, instrument-free, and low-input prediction mechanism for the Shape Index (SI) is still lacking (2, 5). Numerical estimation of SI often requires measurement infrastructure, time, and labor, while the critical point in production lines is a rapid and highly reliable separation of outliers (very

low/very high SI) (16, 20). Although the use of eggshell color as a determinant for SI may seem controversial at first glance, the literature has shown that this variable is not a purely coincidental phenotype, but is associated with underlying biological processes. The shell color and the mottling pattern are not merely visual features, but a result of pigmentation processes (specifically protoporphyrin IX and biliverdin deposition) occurring in the shell gland (25). This pigmentation process is closely related to mineralization, calcification, and shell matrix organization that occur during shell formation (14, 23). Furthermore, the absence of a universal SI threshold standard in the literature necessitates flexible and adjustable decision policies sensitive to different line/age/environmental conditions. In this context, an approach that defers measurement to a secondary stage and generates calibrated probabilities based solely on easily observable phenotypes (weight and shell color), while explicitly managing the error-coverage trade-off, offers both operational feasibility and methodological clarity.

This study aims to investigate the predictability of SI in quail eggs without field measurement (non-invasive, instrument-free) and to develop and validate an automatic decision policy with a rejection option based solely on weight and shell ground color inputs. Given the absence of an operational threshold standard for SI in the literature, our goal is:

(i) to define a target space that characterizes the lower and upper tails of the distribution,

(ii) to establish a two-threshold accept/reject rule using a model that generates calibrated probabilities in this space, and

(iii) to quantitatively optimize the rule's coverage–accuracy tradeoff using a leak-free (out-of-fold, OOF) threshold adjustment.

The outputs of the study aim to provide interpretable, data-driven decision strips for the automatic classification of "clear" cases and the referral of uncertain samples for measurement in field applications. Against this background, our aim is not to regress the continuous shape index directly – an approach that demands full measurements and exhibits low explanatory power with only two field-level inputs – but rather to separate the practically critical extremes (LOW/HIGH SI) with calibrated confidence and a reject option. By tuning a two-threshold policy on out-of-fold probabilities and snapping to the final model's discrete support, we explicitly target the risk–coverage trade-off relevant to hatchery and quality-control settings: confident acceptance of clear cases while routing ambiguous eggs to measurement.

This study differs from the existing approaches by focusing on reliably discriminating application-critical SI values in the lower and upper tails of the distribution, rather than directly regressing the shape index, a continuous variable. Direct regression-based methods generally have limitations, such as the need for precise measurements and low explanatory power under limited field inputs. In contrast, the proposed approach explicitly models decision reliability through calibrated probabilities and a rejection option, and incorporates uncertainty management into the process.

Material and methods

The study material consisted of quail eggs from the Alternative Poultry Breeding Unit at the Experimental Research Application and Research Center of Hatay Mustafa Kemal University. At the start of the study, the quails were 70 days old, and the study lasted 10 weeks. The eggs were collected daily and visually classified as brown or grayish white based on the shell background color (Fig. 1). Each classified egg was individually weighed with a digital scale accurate to 0.01 g, and the egg length and width were measured with digital calipers. The egg shape index (%) was calculated by dividing the egg width by the egg length and multiplying by 100.

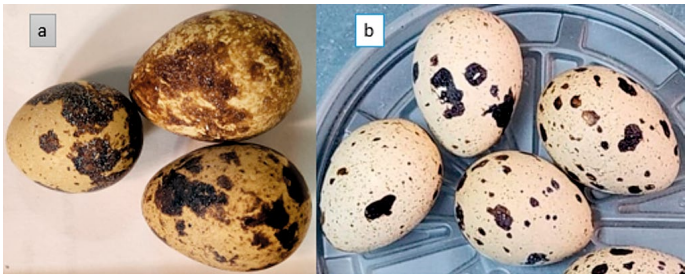

Fig. 1. Quail eggshell ground color: a) Brown ground, b) Grayish white ground

Classification theory with the rejection option. In a classification problem with two classes $C = \{C_1, C_2\}$, x $\in R^p$ characterized by the feature vector and $y \in C$ by the label. The posterior probability is defined by Bayes's formula:

$$p(C_i|x) = \frac{p(x|C_i)p(C_i)}{p(x)} = \frac{p(x|C_i)p(C_i)}{\sum_{i=1}^{2} p(x|C_i)p(C_i)}$$

Here, $p(C_i)$ is the prior probability of class C_i, $p(x|C_i)$ is the conditional probability of x given C_i, and $p(x)$ is the probability of x. There exists a function f: $R^p \rightarrow C$ that divides into two regions (R_1, R_2), with one region for each predicted class, such that $x \in R_i$ implies $f(x) = C_i$. The performance of a classifier is measured by the error rate given below:

$$\varepsilon[f] = p(f(x) \neq y) = \sum_{i=1}^{2} \int_{R_i} \sum_{j=1:j\neq i}^{2} p(x|C_j)p(C_j)\, d_x$$

The accuracy of the classifier is given by $a[f] = 1 - \varepsilon[f]$. The classifier that minimizes error is called the Bayes classifier. This classifier predicts the class with the highest posterior probability:

$$f_{Bayes}(x) = \arg\max_{C_i}(p(C_i|x))$$

If the accuracy of the Bayesian classifier is insufficient for the task at hand, instead of classifying all examples, an approach can be adopted that classifies only those with sufficiently high posterior probabilities. Based on this principle, Chow (1970) presented an optimal classifier with a rejection option (4). A rejection region R_{reject} is defined in the feature space, and all examples belonging to this region are rejected

by the classifier. An example x is accepted only if the probability that x belongs to C_i is equal to or higher than a given probability threshold t:

$$f(x) = \begin{cases} arg\max_{C_i}\big(p(C_i|x)\big) if \ \max_{C_i}(p(C_i|x)) \geq t \\ reject \ if \ p(C_i|x) < t \ \forall i \end{cases}$$

Rejection rate,

$$p(reject) = \int_{R_{reject}} p(x)dx = p(\max\,(p(C_i|x)) \leq t)$$

and the acceptance rate is $p(accept) = 1 - p(reject)$.

There is a general relationship between error and rejection rates: as the error rate decreases monotonically, the rejection rate increases (4). Based on this relationship, Chow proposes an optimal balance between error and rejection.

In Chow's theory, an optimal classifier can only be found if the true posterior probabilities are known. In practice, this is rarely the case. Fumera et al. (2004) show that if there is a significant error in probability estimation, Chow's rule does not perform well (8). In this case, they claim that defining different thresholds for each class yields better results. The classification rule is as follows:

$$f(x) = \begin{cases} arg\max_{C_i}\big(p(C_i|x)\big) if \ \max_{C_i}(p(C_i|x)) \geq t_i \\ reject \ if \ p(C_i|x) < t_i \ \forall i \end{cases}$$

These types of classifiers are popular in the machine learning community. Keep in mind that this method is similar to the concept of soft classification. The main difference is that in soft classification, the classifier's output are *a posteriori* probabilities. In classification with a rejection option, the decision is made based on these posterior probabilities. The classifier's output is either a class assignment or a rejection (11).

Discriminant-based decision rule and rejection. In a two-class classification, the classifier f: $\mathbb{R}^p \rightarrow \{0,1\}$ can be equivalently defined as a discriminant function d: $\mathbb{R}^p \rightarrow \mathbb{R}$, and a class decision can be made based on its sign. Simple (non-rejection) rule:

$$f(x) = \begin{matrix} 0, & d(x) \leq 0 \\ 1, & d(x) > 0 \end{matrix}$$

and the magnitude of $|d(x)|$ reflects the confidence of the decision (distance from the boundary).

The decision is divided into three regions using two thresholds (t_1, t_2) with the rejection option:

$$\delta(x) = \begin{cases} 0, & d(x) \leq t_1 (LOW \ accept) \\ 1, & d(x) \geq t_2 (HIGH \ accep) \\ \emptyset, & t_1 < d(x) < t_2 \ (reject, unsure) \end{cases}$$

Here $t_1 < t_2$, and the rejection region is (t_1, t_2) (12).

If we denote the class-conditional densities for $S = d(X)$ by $g_i(s) = p(S = s|y = i)$, then the class-conditional errors in the accepted cases are

$$\varepsilon_2(t_1) = \mathbb{P}(\delta(X) = 0|y = 1) =$$

$$= \int_{-\infty}^{t_1} g_1(s)ds \ , \varepsilon_1(t_2) = \mathbb{P}(\delta(X) = 1|y = 0)$$

$$= \int_{t_2}^{+\infty} g_0(s)ds$$

The total error $\varepsilon(t_1,t_2) = \pi_1\varepsilon_2\,(t_1) + \pi_0\varepsilon_1(t_2)$. Rejection and coverage:

$$\mathbb{P}(reject) = \sum_{i=0}^{1} \pi_i \int_{t_1}^{t_2} g_i(s)ds, \quad Coverage = 1 - \mathbb{P}(reject)$$

Sensitivity and balanced accuracy:

$$TPR = \int_{t_2}^{+\infty} g_1(s)ds \ , TNR = \int_{-\infty}^{t_1} g_0(s)ds \ , BA = \frac{1}{2}(TPR + TNR)$$

which follows the classical rejection option theory of Chow and subsequent generalizations in statistical pattern recognition and modern learning theory (6, 13).

OOF threshold setting and "snap" procedure. In a classifier with rejection options, the fundamental parameters are the threshold values that define areas to be rejected. Various strategies have been proposed to find the optimal rejection rule. Landgrebe et al. (2006) define 3D ROC curves for a classifier, where the axes represent the true positive rate, the false positive rate rejected by the classifier, and the false positive rate accepted by the classifier (18). The optimal thresholds are selected by maximizing the volume under the 2D surface. Dubuisson and Masson (1993) propose a rejection rule for problems where classes are poorly known (6). It includes two rejection options: uncertainty rejection when an example lies in the area between several classes, and distance rejection for examples far from known class examples. Li and Sethi (2006) suggest controlling error rather than finding a balance between rejection and error rates (19). They frame the problem as designing a classifier with the smallest rejection rate when an error rate is given for each class. Our approach proposes an out-of-fold (OOF) threshold setting. This approach involves three steps:

1. OOF probabilities: With K-fold stratified CV, each sample receives $\hat{p} = \hat{p}(HIGH|x)$ from a model it has not seen. Thus, the probability distribution is leak-free (17, 27).

2. Coverage target: For the desired acceptance rate (coverage) α, p_{low} and p_{hi} are adjusted in small steps on the OOF distribution. Constraints, such as minimum bandwidth and minimum acceptance count in both classes, are applied (3, 13).

3. Final fit + discrete snap: The final model is rebuilt with all training data. Thresholds are iteratively and stably adjusted by "snapping" to the model's actual discrete $\hat{p}$ support points (10, 24).

Final fit and discrete "snap". After OOF tuning, the model was refit on the full tail data, and final probabilities $\{\hat{p}i\}$ were obtained. Because tree-based probability outputs are discrete on a finite support, we snapped (p_{lo}, p_{hi}) to the nearest attainable probability grid points of the final model. A two-phase refinement ensured that (i) both acceptance sides are active and (ii) realized coverage lies within a slightly narrower operational envelope (default [0.45, 0.70]). This yields stable, reproducible operating points in deployment.

Performance metrics and uncertainty. Primary operating metrics were computed on the accepted subset (instances with a definite LOW/HIGH decision):

• Balanced accuracy (BA), accuracy (ACC), sensitivity (recall for HIGH), and specificity (recall for LOW).

• Coverage within extremes: fraction of the samples in the tails that are accepted.

• Expected overall automation: coverage × (extreme-set share in the full data).

Uncertainty was quantified by nonparametric percentile bootstrap (B = 2,000 resamples) on the accepted subset, reporting 95% CIs for BA, ACC, sensitivity, specificity, and coverage.

Probability quality (calibration). We assessed calibration on the accepted subset via

• Brier score,

• Expected Calibration Error (ECE) using 10 equal-width bins on [0,1].

ECE and Brier were also reported by color subgroup (brown vs. grayish white) to probe group-level probability fidelity.

Computational environment. All implementations were carried out with Python 3.12.9. Deep learning models were implemented with TensorFlow 2.19.0 and Keras 3.9.0, while classical machine learning utilities were employed from Scikit-learn 1.6.1. All experiments were conducted in a CPU-only environment because of hardware constraints, though the proposed pipeline is fully compatible with GPU acceleration.

Results and discussion

The study was conducted on a dataset consisting of a total of 1,031 eggs. The distribution of measurements was as follows: weight (g) average 11.879 ± 1.395, length (mm) average 33.194 ± 1.736, width (mm) average 23.847 ± 1.902, and shape index (SI,%) average 71.960 ± 5.963. The eggshell background color distribution was determined as brown 638 (61.9%) and grayish white 393 (37.6%).

Preliminary assessment of SI regression without measurements. The performance of models established for numerical SI estimation using only weight + color inputs is low: Ridge $R^2 \approx 0.018$ and Hist Gradient Boosting $R^2 \approx -0.033$. In scenarios where SI is divided into 3-4 classes, BA remained at ≈ 0.38 and ≈ 0.30. These findings indicate that the two-variable input is insufficient to explain SI variance and that the direct regression approach has limited operational utility. Therefore, the analysis strategy focused on binary classification with rejection options.

Definition of the tail regions and the analysis set. With this approach, the sample quantiles for SI were calculated as $t_1 = 67.197$ and $t_2 = 76.196$. The lower and upper tails defined by these thresholds ($SI < t1 \vee SI \geq t2$) contain 547 (53.1%) observations, and the automatic decision policy was applied only to this subset.

Modeling: calibrated probabilities and two-threshold decision. A classifier calibrated using only weight and color inputs was created in the extreme tails (Hist Gradient Boosting + target encoding + Platt "sigmoid"). The model produces $\hat{p} = P(\text{HIGH} \mid x)$ for each example. The decision rule is defined as LOW if $\hat{p} < p_{lo}$, HIGH if $\hat{p} > p_{hi}$, and UNSURE (referral for measurement) otherwise. The thresholds were selected with the coverage target on the leak-free OOF probability distribution and stabilized by "snapping" to the final model's discrete probability support. At the final working point, $p_{lo} = 0.5274$ and $p_{hi} = 0.7609$ were obtained.

Operational performance: coverage and accuracy. Performance on accepted cases: coverage in the extreme tails = 0.614; BA = 0.720, sensitivity (HIGH) = 0.672, specificity (LOW) = 0.767, accuracy 0.717. The confusion matrix was found to be TN = 122, FP = 37, FN = 58, TP = 119. On accepted samples, 2,000 repeated bootstraps yielded BA point estimate 0.72 with 95% CI: [0.67, 0.77], ACC 0.72 with 95% CI: [0.67, 0.76]. Probability quality was found to be Brier = 0.227 and ECE (10-thousand) = 0.064. This indicates that the calibration curve is generally well-aligned, but there is a small margin for improvement in the middle probability bands. Considering the proportion of the tail set in the entire data (53.1%), the expected total automation rate is approximately 0.614 × 0.531 ≈ 32.6% (Tab. 1).

Tab. 1. Summary metrics of the final two-threshold policy (extreme tails)

Metric	Value
Lower/upper SI quantiles (t_1, t_2)	67.197; 76.196
Probability thresholds (p_{lo}, p_{hi})	0.5274; 0.7609
Coverage within extremes	0.6143 (≈ 61.4%)
Balanced accuracy (BA)*	0.720 (95% CI: 0.670-0.770)
Sensitivity – HIGH	0.672
Specificity – LOW	0.767
Accuracy (accepted only)	0.717 (95% CI: 0.670-0.760)
Brier score (accepted only)	0.227
ECE (10-bin, accepted only)	0.064
Confusion matrix (accepted only)	TN = 122, FP = 37, FN = 58, TP = 119
Extremes share (of all data)	0.5306 (547/1031)
Expected overall automation†	0.3259 (≈ 32.6%)

Explanations: * – 95% CIs for BA and ACC computed with 2,000 bootstrap resamples (within the accepted subset). † Expected overall automation = (coverage within extremes) × (extremes share) = 0.6143 × 0.5306 ≈ 0.3259. Decision rule: accept LOW if wzór; accept HIGH if wzór; otherwise UNSURE (send to measurement). Probabilities are Platt-calibrated ("sigmoid"); thresholds are tuned on OOF probabilities for target coverage and "snapped" to the final model's discrete support.

Risk–coverage envelope and selection of the working point. The grid scan conducted with OOF probabilities produced a risk–coverage envelope (Fig. 2a). As coverage increased, the accepted values rose from BA ~0.50 to ~0.56, reaching saturation at a coverage level of ≈ 0.80. The selected operating point (Fig. 2b) is positioned within the target coverage band. After final calibration and "snap," BA ≈ 0.72 and coverage ≈ 0.61 were achieved.

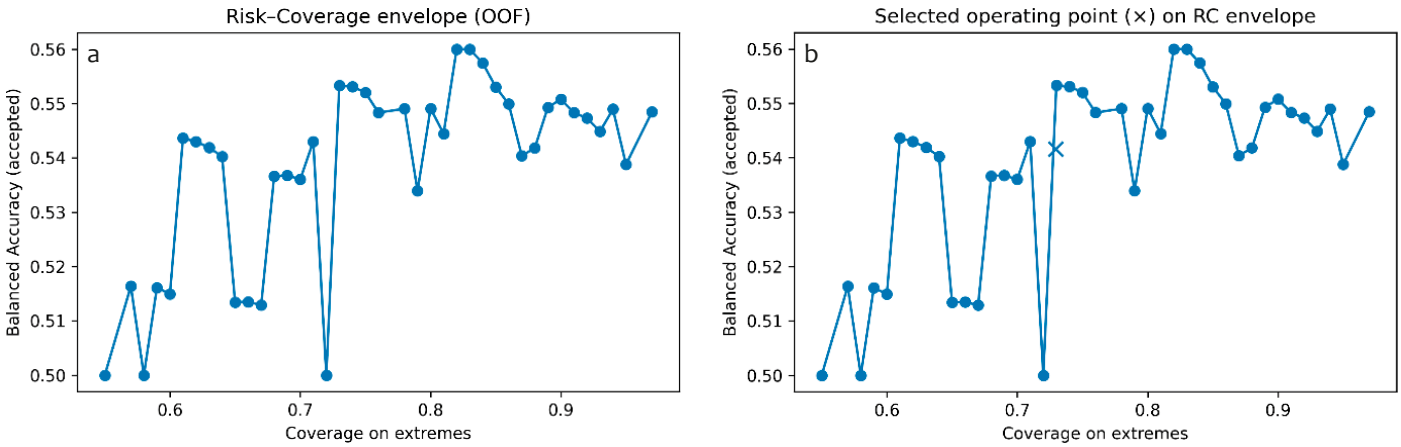

Fig. 2. OOF risk–coverage envelope and the selected operating point. (a) Risk–coverage envelope derived from out-of-fold probabilities. (b) The chosen operating point (×) on the envelope indicating the final threshold pair and expected coverage

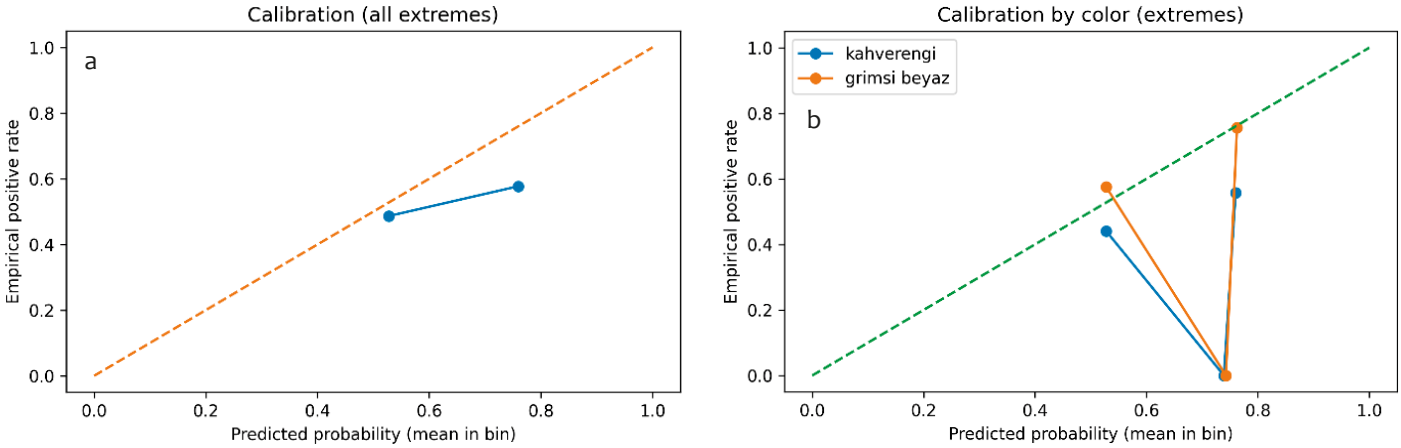

Fig. 3. Calibration curve. (a) Calibration curve for extreme tails (orange: ideal $y = x$). (b) Calibration curves by color

Probability quality and uncertainty structure. In the accepted samples, the Brier score = 0.227 and ECE (10-k) = 0.064 were calculated. These values indicate that the calibration curve is generally well-aligned but shows slight overconfidence in the medium probability bands (Tab. 1). Using bootstrap (n = 2,000) on the accepted cases, 95% CI: BA [0.67-0.77], ACC [0.67-0.76] were obtained.

Calibration (general and by color). The calibration curve in the extreme tails (Fig. 3a) consists of two points due to the bimodal distribution, and both points lie below the ideal diagonal, indicating slight overconfidence. When examining color-based calibration (Fig. 3b), overconfidence is more pronounced in the brown group, while calibration is close to the diagonal for the high probability mode in the grayish white group. Thanks to the OOF-coverage-based selection of thresholds, these calibration deviations do not adversely affect operational metrics.

Operational decision bands. The decision bands derived in the weight-color plane (Fig. 4) show that HIGH decision regions dominate at low weights and LOW decision regions dominate at high weights,

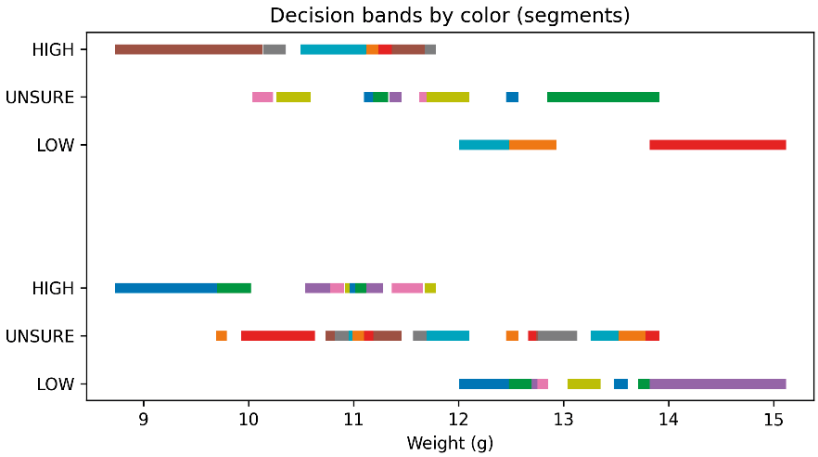

Fig. 4. Weight–decision bands by color

whereas UNSURE windows specific to color appear at medium weights. The approximate intersections of the threshold probabilities on the weight axis are $w_{p_{lo}} \approx 12.06$ g and $w_{p_{hi}} \approx 11.74$ g for both colors.

In this study, we presented a classification framework with probability calibration and a rejection option aimed at automatically separating application-relevant extremes (low/high) in the tails of SI in quail eggs using only weight + shell color information. Unlike the classical approach of direct regression, we focused on

selecting operationally critical extreme cases with high confidence. This provides a practical pre-screening layer that reduces the measurement burden in incubation and quality control lines. Our findings show that SI outliers can be separated even with a low input set, while rejected "uncertain" samples are rationally separated for measurement/second-stage review. This design makes it possible to integrate two-threshold policies that explicitly manage the error–coverage tradeoff into industrial decision processes. The theoretical underpinnings of this framework are Chow's rule, which optimizes the error–rejection tradeoff, and the subsequent rejection-option classification literature (4).

The classical literature shows that SI can establish meaningful relationships with quality indicators, such as specific gravity, albumen index, and Haugh unit, but the effects are sensitive to strain/age/environment (7). Recent studies have reconfirmed that the shell color/pattern can play a discriminating role in the internal/external quality and hatchability of quail and chicken eggs. For example, Ismael et al. (2024) found color-based differences in fertility and hatchability rates in a large series comparing five color/pattern types, while reporting that SI was not significantly affected by color (15). Species/line-specific current data, however, show an interaction of color diversity with hatchability and quality metrics; when weight, width, and shell quality are used together, the discriminatory power increases (9, 21). Consistent with these findings, our approach aims to reliably separate operationally critical edge classes using calibrated probabilities and two thresholds, rather than approximating SI directly through regression, thus providing a practical and reproducible method for unmeasured pre-screening in production/hatching lines.

The rejection option literature has shown that single-threshold and class-dependent multi-threshold rules can improve the error–rejection and cost tradeoff. In particular, work along the lines of Fumera et al. demonstrates that defining different thresholds per class yields better results when calibration errors are high. More recent syntheses (e.g., the "reject option" survey) distinguish between uncertainty and novelty rejection, proposing threshold optimization through ROC surfaces and volume metrics. Our two-threshold policy design (LOW/HIGH acceptance, neutral in the intermediate zone) is an operational adaptation of this line: it manages quality/risk by controlling the acceptance rate at the expense of reducing the expected error for each egg accepted on the production line (8, 26).

The study has some limitations: (i) The model uses only two attributes (weight and color); this deliberate simplification increases operational applicability, while the addition of image-based phenotypes, such as pigment area/brightness intensity/staining percentage, has the potential to shift the coverage–accuracy curve upward. (ii) The data may be single-period/single-herd; multi-center and time-spanning sampling would strengthen generalizability. (iii) Our decision not to perform direct regression toward SI is a methodological choice; the goal of reliably separating outlier classes aligns better with the requirement for low-error-cost, coverage-controlled decisions on production lines.

Future work on integrating shell phenotypes extracted from images (pigmentation intensity, texture/texture metrics), line/age-stratified models, cost-sensitive coverage optimization, prospective field validation, and online band adaptation are natural extensions to improve the method's accuracy and coverage. Additionally, multi-objective selection schemes with quality-hatch outputs (e.g., hatch success) will enable dynamic policy adjustments based on production targets.

Using only egg weight and shell color, a two-threshold policy based on calibration probabilities for samples in the tails of the SI distribution achieved reliable accuracy at BA ≈ 0.72 and ≈ 32-33% total automation. The policy was calibrated with OOF-based and coverage-targeted threshold adjustment; it was made operational with color-based calibration and decision bands. This approach rationalizes resource usage by automatically classifying "clear" cases while directing uncertain (UNSURE) cases to measurement.

References

1. *Alaşahan S., Çopur Akpınar S., Canoğulları S., Baylan M.*: Determination of some external and internal quality traits of Japanese quail (Coturnix coturnix japonica) eggs on the basis of eggshell colour and spot colour. Eurasian J. Vet. Sci. 2015, 31 (4), 235-241, doi: 10.15312/EurasianJVetSci.2015413529.
2. *Assefa S., Abebe B. K., Gobena A. H.*: A study on egg quality and hatching traits of indigenous and exotic chickens reared in Silte zone, Southern Ethiopia. Heliyon 2023, 9 (8), e19126, doi: 10.1016/j.heliyon.2023.e19126.
3. *Bartlett P. L., Wegkamp M.*: Classification with a reject option using a hinge loss. Journal of Machine Learning Research 2008, 9, 1823-1840, doi: 10.1145/1390681.1442792.
4. *Chow C. K.*: On optimum recognition error and reject tradeoff. IEEE Trans. Inf. Theory 1970, 16, 41-46, doi: 10.1109/TIT.1970.1054406.
5. *Drabik K., Batkowska J., Vasiukov K., Pluta A.*: The impact of eggshell colour on the quality of table and hatching eggs derived from Japanese quail. Animals 2020, 10 (2), 264, doi: 10.3390/ani10020264.
6. *Dubuisson B., Masson, M.*: A statistical decision rule with incomplete knowledge about classes. Pattern Recognition 1993, 26 (1), 155-165, doi: 10.1016/0031-3203(93)90097-G.
7. *Duman M., İekeroğlu A., Yıldırım A., Eleroğlu H., Camcı Ö.*: Relation between egg shape index and egg quality characteristics Zusammenhang zwischen Formindex des Eies und Eiqualitätsmerkmalen. European Poultry Science 2016, 80, 1-9, doi: 10.1399/eps.2016.117.
8. *Fumera G., Pillai I., Roli F.*: A two-stage classifier with reject option for text categorisation. Conference: Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR International Workshops, SSPR 2004 and SPR, Lisbon, Portugal 2004, 18-20, doi: 10.1007/978-3-540-27868-9_84.
9. *Gu H., Yan Z., Zhang B., Chen X., Geng A., Zhang Y., Cao J., Zhang J., Zeng L., Wang Z., Liu H., Chu Q.*: Impact of eggshell color diversity on hatchability, translucency, and quality traits in Beijing-You chicken eggs. Animals (Basel) 2025, 15 (17), 2595, doi: 10.3390/ani15172595.
10. *Guo C., Pleiss G., Sun Y., Weinberger K. Q.*: On calibration of modern neural networks. arXiv 2017, 1706.04599, doi: 10.48550/arXiv.1706.04599.
11. *Hanczar B., Dougherty E. R.*: Classification with reject option in gene expression data. Bioinformatics 2008, 24 (17), 1889-1895, doi: 10.1093/bioinformatics/btn349.
12. *Hendrickx K., Perini L., Plas Van der D., Meert W., Davis J.*: Machine learning with a reject option: A survey. Mach. Learn. 2024, 113 (5), 3073-3110, doi:10.1007/s10994-024-06534-x.
13. *Herbei R., Wegkamp M. H.*: Classification with reject option. Canadian Journal of Statistics 2006, 34 (4), 709-721, doi: 10.1002/cjs.5550340410.

14. *Hincke M. T., Nys Y., Gautron J., Mann K., Rodriguez-Navarro A. B., McKee M. D.*: The eggshell: structure, composition and mineralization. Front Biosci. 2012, 17 (4), 1266-1280, doi: 10.2741/3985.
15. *Ismael N. A., Abdelmonem U. M., El-Kholy M. S., El Nagar A. G., Ahmed A. F., Almalki M., El-Tarabily K. A., Reda F. M.*: The relationship between eggshell color, hatching traits, fertility, mortality, and some qualitative aspects of Japanese quail (Coturnix japonica) eggs. Poultry Science 2024, 103 (2), 103298, doi: 10.1016/j.psj.2023.103298.
16. *Kruenti F., Lamptey V. K., Okai M. A., Adu-Aboagye G.*: The influence of flock age and egg size on egg shape index, hatchability and growth of Japanese quail chicks. Journal of Innovative Agriculture 2022, 9 (1), 8, doi: 10.37446/jinagri/rsa/9.1.2022.8-16.
17. *Laan M. J. Van der, Polley E. C., Hubbard A. E.*: Super learner. Stat. Appl. Genet. Mol. Biol. 2007, 6, 25, doi: 10.2202/1544-6115.1309.
18. *Landgrebe T. C., Tax D. M., Paclík P., Duin R. P.*: The interaction between classification and reject performance for distance-based reject-option classifiers. Pattern Recognition Letters 2006, 27 (8), 908-917, doi: 10.1016/j.patrec.2005.10.015.
19. *Li M., Sethi I. K.*: Confidence-based classifier design. Pattern Recognition 2006, 39 (7), 1230-1240, doi: 10.1016/j.patcog.2006.01.010.
20. *Li S., Zhao X., Zheng X., Chen H., Zhou R., Shi L., Liu H., Xu L., Ning Z., Wang D.*: Study on changes in egg quality traits and genetic parameters of white leghorn hens from 35 to 100 weeks of age. Poult. Sci. 2025, 104 (10), 105502, doi: 10.1016/j.psj.2025.105502.
21. *Ludoški M., Grkovic N., Suvajdzic B., Vicic I., Lazic I. B., Baltic T., Cobanovic N.*: Eggshell colour affecting the safety and quality of Japanese quail eggs (Coturnix coturnix japonica). Br. Poult. Sci. 2025, 1-13, doi: 10.1080/00071668.2025.2527225.
22. *Mizutani M.*: The Japanese quail. Laboratory Animal Research Station, Nippon Institute for Biological Science, Kobuchizawa, Yamanashi, Japan 2003, 143-163.
23. *Nys Y., Gautron J., Garcia-Ruiz J. M., Hincke M. T.:* Avian eggshell mineralization: biochemical and functional characterization of matrix proteins. Comptes Rendus Palevol 2004, 3 (6-7), 549-562, doi: 10.1016/j.crpv.2004.08.002.
24. *Niculescu-Mizil A., Caruana R.*: Predicting good probabilities with supervised learning. ICML '05: Proceedings of the 22nd International Conference on Machine Learning 2005, 625-632.
25. *Samiullah S., Roberts J. R., Chousalkar K.*: Eggshell color in brown-egg laying hens – a review. Poultry Science 2015, 94 (10), 2566-2575, doi: 10.3382/ps/pev202.
26. *Uyar A., Atilgan Sengül Y.*: Rejection threshold optimization using 3D ROC curves: Novel findings on biomedical datasets. International Journal of Intelligent Systems and Applications in Engineering 2021, 9 (1), 22-27, doi: 10.18201/ijisae.2021167933.
27. *Wolpert D. H.*: Stacked generalization. Neural Networks 1992, 5 (2), 241-259, doi: 10.1016/S0893-6080(05)80023-1.

Corresponding author: Betül Dağoğlu Hark, Fırat University, Faculty of Medicine, Department of Biostatistics, Elazığ, Türkiye; e-mail: bdagoglu@firat.edu.tr