**Proposed algorithm**

 kNN categorization (k=5 default):

* safe: ≤1 opposite-class neighbor
* border/overlap: 2–3 opposite neighbors
* outlier: 4–5 opposite neighbors

 Oversampling:

* Use only **safe + border** as seeds (exclude outliers).
* Optional distance-aware refinements (toggleable later):
  + margin-based difficulty (same vs opp mean distance),
  + density-adjusted impurity (impurity × sparsity via r\_k),
  + distance-weighted impurity for ranking/pruning.

 Undersampling (replace RUS/NCL):

* **Redundancy-aware majority pruning**: remove only deep-interior majority points that are same-class surrounded and far from minority (to minimize info loss). Parameters (β target IR reduction, k, density/nearest-minority thresholds).

**Oversampling side (safe/border/outlier idea)**

* **Borderline-SMOTE (Han et al., 2005):** oversamples only borderline minority instances (based on kNN impurity). Very close to your “safe+border, skip outliers” idea.
* **Safe-Level-SMOTE (Bunkhumpornpat et al., 2009):** uses a “safe level” ratio (minority vs majority neighbors) to guide oversampling. If the safe level is low (outlier), no samples are generated. This aligns almost exactly with your intuition about excluding outliers.
* **ADASYN (He et al., 2008):** adaptively generates more synthetic points for harder (high-impurity) minority samples. Different emphasis — it *focuses* on borderline/outliers, whereas you *avoid* outliers.
* **MDO (Majority Directed Over-sampling, 2017):** considers neighborhood structure, but again usually includes borderline+outliers.

**Undersampling side (your iterative redundancy pruning)**

Most existing US methods fall into 2 families:

* **Random-based:** RUS, Cluster Centroids, NearMiss. → All risk high information loss.
* **Cleaning-based:** Tomek Links, Edited Nearest Neighbors (ENN), Neighborhood Cleaning Rule (NCL). → These remove noisy or borderline *majority*, not the deep interior ones.

Closest existing:

* **Condensed Nearest Neighbor (CNN, Hart 1968):** keeps boundary points, discards interior. But CNN is one-pass and aggressive, not iterative.
* **One-Sided Selection (OSS, Kubat & Matwin 1997):** Tomek links + CNN → still boundary-focused.
* **Prototype Reduction (various, 2000s):** cluster-based pruning of majority to reduce redundancy, but usually non-iterative.
* **Neighborhood Tomek (various extensions):** removes pairs, not redundancy-driven.
* **Iterative editing/pruning** methods do exist, but they usually focus on cleaning noisy *minority*, not gradually thinning majority interiors.