# Extended Literature Review: VAE‑based Oversampling for Imbalanced Data

## 1. Over‑Sampling Algorithm Based on VAE (Zhang et al., 2018)

Zhang and colleagues deliver the first fully VAE‑driven oversampler, positioning the model as a distribution‑aware alternative to SMOTE. They note that most resampling methods work either by raw duplication or local interpolation, neither of which respects the \*\*global probability density\*\* of the minority class. Their algorithm therefore trains a vanilla VAE \*solely\* on minority instances, leveraging the encoder–decoder pair to fit an implicit density without any parametric prior assumptions about feature dependence. A practical twist is the separation of \*\*discrete and continuous attributes\*\*: continuous features are generated through the decoder while discrete ones are taken from the nearest real minority neighbour so that synthetic samples stay semantically valid. Experiments on five UCI datasets show that the VAE yields higher F1‑minority and G‑mean than both normal‑distribution oversampling and SMOTE at equal expansion rates; Table 3 in the paper reports up to \*\*3 pp F1 gains\*\* at 300 % oversampling and, importantly, no degradation in majority‑class accuracyfileciteturn1file15. The study concludes that density‑aware generation is a stronger signal than mere sample size for classifier generalisationfileciteturn1file18.

## 2. Generative Oversampling with a Contrastive VAE (Dai et al., 2019)

Dai et al. observe that rare‑event prediction often contains \*\*variance common to both majority and minority cohorts\*\*—for example, baseline vitals in a hospital ward—plus \*class‑specific\* deviations such as adverse outcomes. The proposed \*\*Contrastive VAE (C‑VAE)\*\* decomposes the latent space into a shared subspace (capturing common structure) and a private subspace (capturing class‑specific structure). During generation the algorithm keeps a drawn shared vector fixed and samples only the private minority latent, thus preserving physiological plausibility while amplifying minority‑class idiosyncrasies. On a real clinical dataset with extreme outcome scarcity, logistic‑regression AUC improves significantly over both standard VAEs and the SWIM oversampler; synthetic‑data experiments further confirm that C‑VAE excels precisely when the minority shares at least some covariance with the majorityfileciteturn1file7. The authors caution, however, that if classes are distributed on disjoint manifolds the contrastive assumption may become harmful.

## 3. Majority‑Guided VAE (Ai et al., 2023)

Facing imbalance ratios above 100 : 1 in vision and tabular tasks, Ai et al. repurpose the abundant majority data to \*guide\* minority generation. Their \*\*MGVAE\*\* introduces a \*\*majority‑conditioned prior\*\* p(z | X⁺) constructed as a mixture of Gaussians centred on majority embeddings. Training unfolds in two stages: (i) \*\*pre‑train\*\* the VAE on majority data until the decoder captures rich semantic variability; (ii) \*\*fine‑tune\*\* on the minority class while applying Elastic‑Weight‑Consolidation to avoid catastrophic forgetting. Algorithm 1 provides a one‑to‑one mapping: pick a majority sample, sample a latent from its conditional prior, decode to a new minority instancefileciteturn1file12. Extensive benchmarks on CIFAR‑LT, ImageNet‑LT and eight tabular sets show MGVAE outperforms prior oversamplers on \*\*macro‑F1 by 4–6 pp\*\* and shrinks minority‑majority feature distance, with ablations confirming that removing the majority prior or EWC collapses generationfileciteturn1file2.

## 4. SMOTE‑CLS: CVAE‑Assisted Noise Filtering (Hong et al., 2024)

Hong et al. reprise the classical ‘SMOTE‑fails‑on‑noise’ narrative and propose \*\*SMOTE‑CLS\*\*, a four‑stage pipeline: (1) \*\*difficulty labelling\*\* via a k‑NN classifier splits samples into ‘easy’ and ‘hard’; (2) train a \*\*class‑conditional VAE\*\* whose latent geometry preserves those difficulty cues; (3) apply \*\*density‑based filtering\*\* in latent space to remove isolated or overlapping minority points; (4) execute standard SMOTE only on the cleansed set and map synthetics back to data space. t‑SNE visualisations show markedly cleaner class separation compared with deep generators and SMOTE‑ENNfileciteturn1file3. Across 20 low‑ and high‑dimensional UCI datasets SMOTE‑CLS lifts AUPRC by an average of 4 pp over DeepSMOTE, while maintaining diversity thanks to a latent space explicitly trained for discriminative structurefileciteturn1file8.

## 5. VAE‑SMOTE Augmented Diffusion for Anomaly Detection (Feng et al., 2025)

Feng et al. extend VAE‑based oversampling to \*\*unsupervised anomaly detection\*\* in tabular data. They first train a VAE on normal instances, embed them, and apply SMOTE in latent space—termed \*\*VSA\*\*—to enlarge the pool of ‘normal’ data while respecting non‑linear feature relationsfileciteturn1file4. A diffusion model is then trained on this augmented set; at inference a \*\*truncated diffusion\*\* schedule speeds reconstruction. The widened gap between low reconstruction error for normals and high error for anomalies yields state‑of‑the‑art AUC‑PR on 18 datasets. Ablation studies confirm that VSA contributes more than plain SMOTE, and the full VAE + diffusion stack (Task C) outperforms the diffusion‑only baseline by ~3.5 pp AUC‑PRfileciteturn1file17. Hyper‑parameter sweeps further show robustness to latent dimension and k‑nearest‑neighbour choicesfileciteturn1file14.

## 6. Cross‑Paper Insights

Together these studies trace an evolution from density‑matching generation (2018) through \*\*contrastive and conditional latents\*\* (2019, 2024) to \*\*majority‑guided priors\*\* (2023) and finally hybridisation with diffusion backbones (2025). Two recurring lessons emerge: (i) \*latent‑space regularisation\*—whether via KL divergence, private‑shared decomposition, or density filtering—guards against unrealistic extrapolations; (ii) \*leveraging majority information\*—either as a prior or as negative evidence—injects diversity that small minority sets cannot provide on their own. Remaining gaps include theoretical guarantees for latent‑space filtering thresholds and the scalability of diffusion hybrids to very high‑cardinality multiclass datasets.

# References

1. [1] C. Zhang \*et al.\*, “Over‑Sampling Algorithm Based on VAE in Imbalanced Classification,” LNCS 10967, pp. 334‑344, 2018.
2. [2] W. Dai \*et al.\*, “Generative Oversampling with a Contrastive VAE,” \*IEEE ICDM Workshops\*, pp. 301‑308, 2019.
3. [3] Q. Ai \*et al.\*, “Generative Oversampling for Imbalanced Data via Majority‑Guided VAE,” \*AISTATS\*, vol. 206, pp. 1‑14, 2023.
4. [4] S. Hong \*et al.\*, “Improving SMOTE via Fusing Conditional VAE for Data‑adaptive Noise Filtering,” arXiv:2405.19757, 2024.
5. [5] L. Feng \*et al.\*, “VAE‑SMOTE Augmented Diffusion for Anomaly Detection,” \*Proc. GAIIS\*, pp. 163‑170, 2025.