WHITEPAPER · 16 PAGES / MARCH 2025

Data-Centric AI Under Extreme Well Scarcity: How 235 Patches Became 75,000 and Why 14 Wells Beat Synthetic Overload

In most proprietary subsurface domains the binding constraint is not compute or architecture — it is wells. You cannot acquire more of them, you cannot buy a public corpus that matches your formation, and the labels you do have were written by a handful of expert interpreters. This whitepaper is a field account of building a borehole-image detector under exactly that constraint: fourteen wells of high-resolution borehole image logs from two different microresistivity imaging tools, from a mid-sized Middle East carbonate operator we partnered with, and a reservoir interval that started life with thirty-two labelled sinusoids. We show how the model quality that mattered was won data-side rather than model-side. Overlapping-window patching plus a six-transform augmentation regimen applied ten times per labelled patch scaled the working sinusoid set roughly sixty-five-fold — from about nine hundred image-and-ground-truth pairs to over fifty-five thousand, and from one interval's two hundred thirty-six patches to four thousand two hundred twelve. Augmentation alone collapsed a classification error from a degenerate one hundred percent to 2.618 percent. Adding real wells, three to fourteen, drove classification error from 93.115 to 2.536 — and crucially, it beat the obvious shortcut of inflating the corpus to ninety-two thousand augmented samples, which made the model worse. The lesson generalises: in scarce proprietary domains, data engineering is the model engineering, and not all data is the kind that helps.

Quamer Nasim
ML Research Engineer

Tarry Singh
Founder & CEO

DATA-CENTRIC-AI / DATA-AUGMENTATION / WELL-SCARCITY / SUBSURFACE-AI / COMPUTER-VISION / DETECTION-TRANSFORMER

CONTENTS

There is a kind of AI problem where the constraint is not the model. You can reach for the best architecture, the largest backbone, the most aggressive optimiser, and none of it will move the number, because the thing starving the model is the thing you cannot order more of: data. In proprietary subsurface domains this is the normal condition, not the edge case. A carbonate operator does not have ten thousand wells with expert-validated borehole interpretations. It has fourteen. There is no public corpus that matches its formation, its tool string, or its interpreters' conventions, and the labels it does hold were produced by a small number of geoscientists whose time is the scarcest resource of all. In that regime the entire game moves data-side. The model engineering becomes data engineering — and the operators who win are the ones who understand that not all data is the kind that helps.

This whitepaper is a field account of building under exactly that constraint. Working with a mid-sized Middle East carbonate operator, we built a computer-vision detector for bedding planes, fractures, and dip-azimuth geometry on high-resolution borehole image logs — captured with two different microresistivity imaging tools — across a multi-phase formation-evaluation engagement. The supervised core was a Detection-Transformer-derived model that frames borehole-feature picking as end-to-end set prediction. But the headline of this piece is not the architecture. It is that the decisive gains came from how we manufactured, conditioned, and grew the training data out of a vanishingly small base — and from a discipline that resisted the single most tempting shortcut in small-data computer vision: dumping synthetic volume into the corpus and hoping scale would substitute for signal.

The starting condition: thirty-two sinusoids and a degenerate model

To make the scarcity concrete, start at the smallest unit of the problem. A single reservoir interval in one well, a depth band of roughly sixty metres, yielded thirty-two labelled sinusoids — the sine-shaped traces a planar feature leaves when the cylindrical borehole wall is unrolled into a flat image. Tiled into fixed-height patches, that interval produced two hundred thirty-six image patches, of which only nineteen contained a sinusoid at all. That is the real shape of proprietary subsurface data: not merely small, but savagely imbalanced, with the signal class appearing in under one patch in twelve.

A detector trained on that distribution does the rational thing and the useless thing simultaneously: it learns to predict the majority class everywhere and is rewarded for it. In our augmentation ablation, a model trained without any augmentation posted a classification error of one hundred percent — a fully degenerate predictor that never correctly committed to the minority class. That number is not a near-miss to be tuned away with a better learning rate. It is the signature of a data problem masquerading as a model problem, and no amount of architecture search fixes it. The fix has to change the data the model sees.

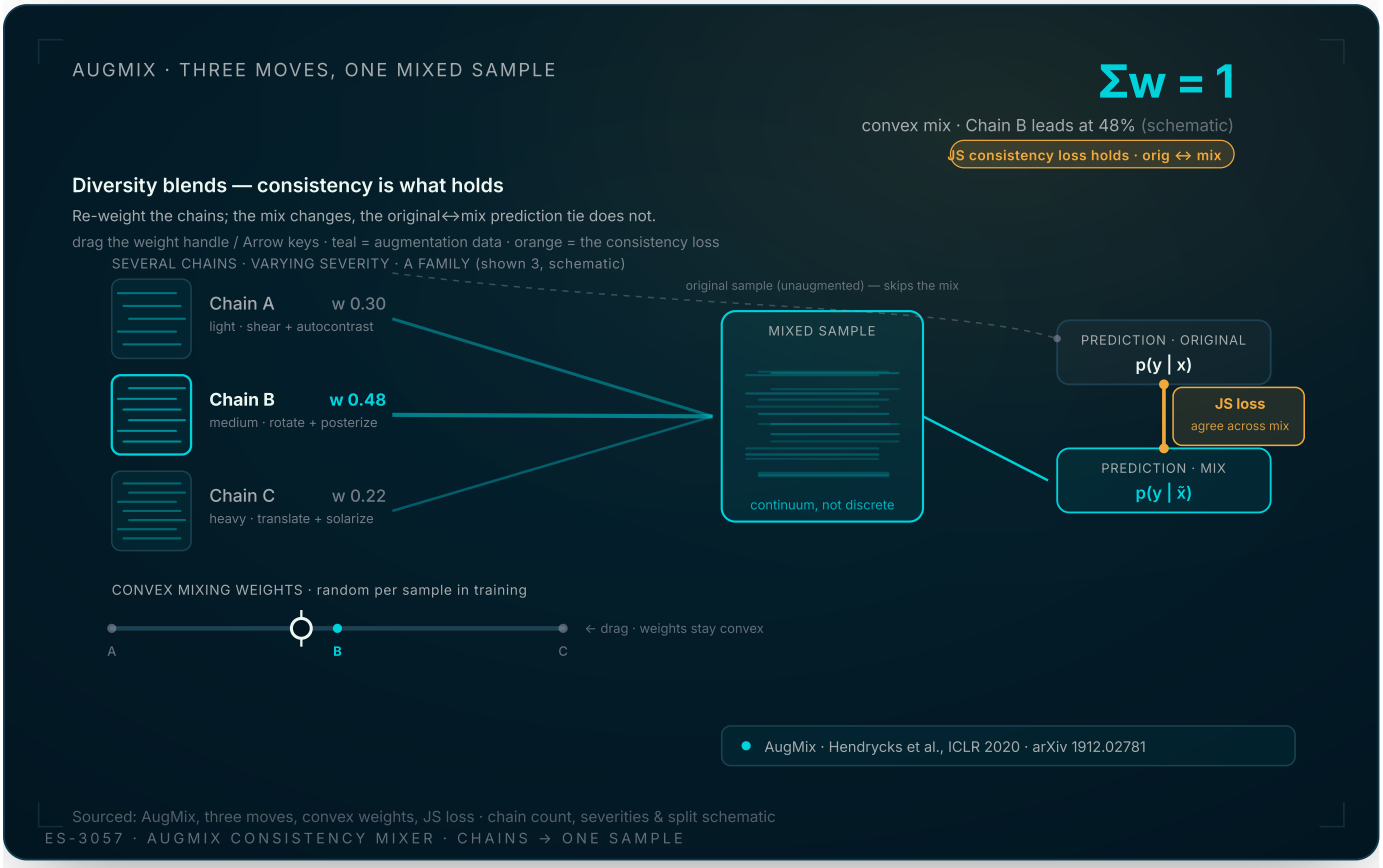
Engineering data out of scarcity: overlapping windows and a six-transform regimen

Two data-side moves did the heavy lifting, and they compound.

The first is **overlapping-window patching**. A raw borehole image is enormous — on the order of hundreds of thousands of pixels tall by three hundred sixty wide, one column per degree around the borehole — and the naive move is to cut it into non-overlapping tiles. That throws away data and, worse, slices features at arbitrary tile boundaries. Instead we generated patches with a sliding window and a deliberately short stride, so consecutive patches overlap heavily. A feature that falls awkwardly at one patch's edge sits cleanly inside the next. The same physical sinusoid is presented to the model many times, at many vertical offsets, each a legitimately distinct training example rather than a duplicate. Stride is a lever here, not a constant: too aggressive and you flood the set with near-identical patches that teach nothing; too conservative and you leave signal on the floor. We tuned it explicitly, and at one point relaxing the stride from forty to eighty pixels was the change that lifted a struggling model — a reminder that the patching schedule is a hyperparameter as load-bearing as anything in the network.

The second move is **augmentation as image-degradation modelling**, not as generic noise. Borehole image logs are not clean photographs; they are physical measurements corrupted in characteristic ways — pad-pressure variation, tool-speed artefacts, resistivity contrast that brightens and dims, blur from the logging run. So we built a six-transform augmentation regimen chosen to mimic exactly those corruptions: colour jitter, Gaussian blur, sharpening, Gaussian noise, emboss, and median blur. Each labelled, sinusoid-bearing patch was passed through this regimen ten times, producing ten variants that preserve the geometry of the feature — its depth, dip, and azimuth are unchanged — while varying the

appearance the way the borehole environment actually varies it. We augmented only the sinusoid-bearing patches, not the whole set, which simultaneously grows the signal class and rebalances the distribution.



The article's most concrete mechanism is AugMix, not a benchmark number. The article describes AugMix as three moves: generate several augmentation chains of varying severity ('a family of them') applied to the same sample, mix their outputs with random convex weights into one continuum-of-corruptions sample, then tie the model's prediction on the original to its prediction on the mix with a Jensen-Shannon-divergence consistency loss. Drag the convex-weight handle: the chains re-weight and the mixed sample re-blends live, but the orange JS-consistency tie holds the original and mixed predictions together no matter how the weights move — that, the article argues, is what makes AugMix robust where naive augmentation bakes in distortion. Sourced from the article: AugMix (Hendrycks et al., ICLR 2020, arXiv 1912.02781), the three moves, 'several chains of varying severity', the random convex mixing weights, and the Jensen-Shannon consistency loss. The article does not fix the number of chains, so the three chains shown (A/B/C), their severities, and the live weight split are schematic and flagged as such on the canvas.

The instrument above illustrates the principle that makes augmentation work rather than merely inflate: a robust detector should predict the same geometry whether it sees a clean patch or a corrupted one, so the augmented variants are not new facts but consistency constraints tying the model's prediction on a degraded image to its prediction on the clean original. Our pipeline is the disciplined, domain-specific instance of that idea — six physically-motivated transforms, ten variants per labelled patch — rather than a generic

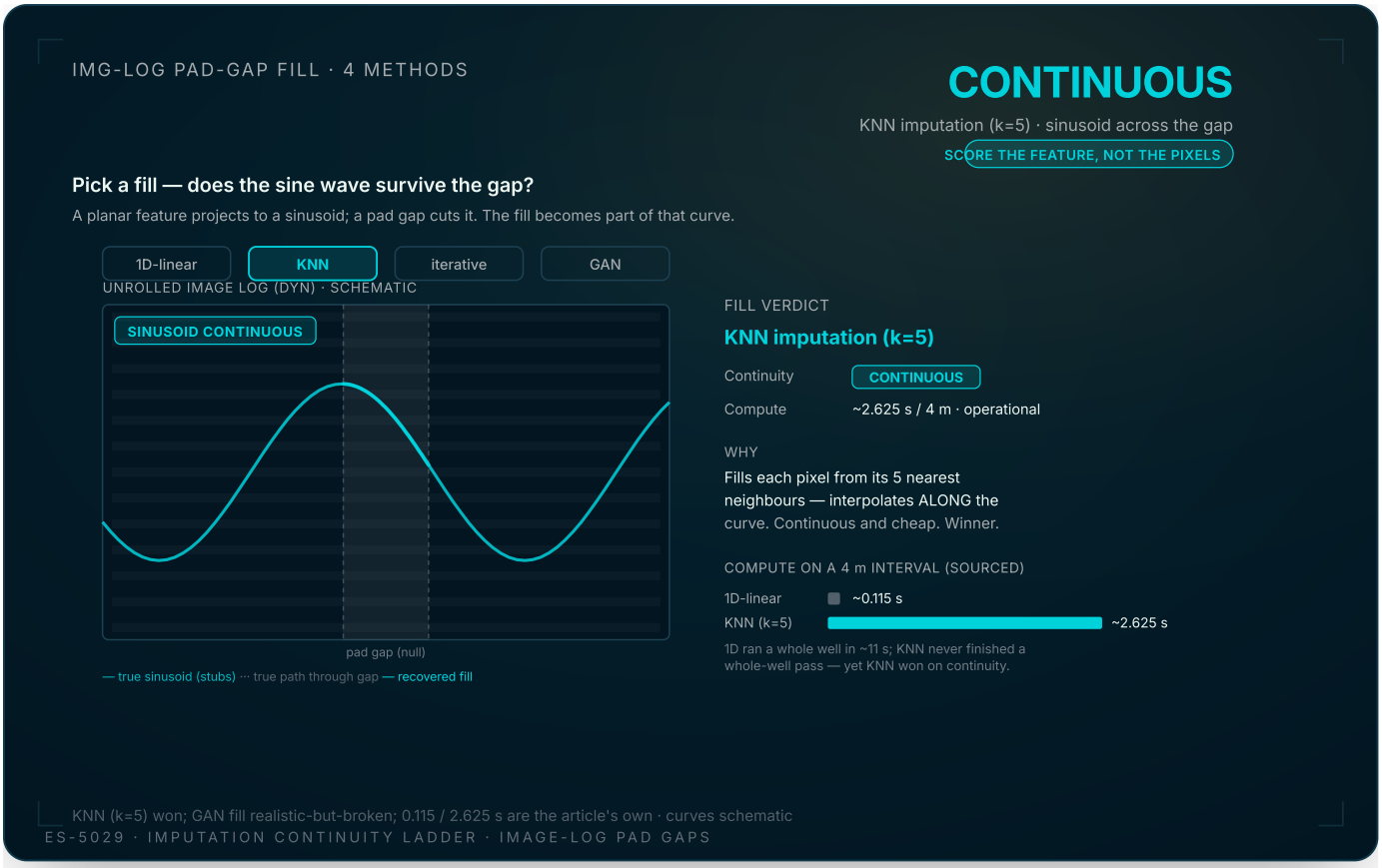
chain. The transforms are chosen because they span the corruption manifold of real high-resolution borehole image logs, which is why a model trained through them holds up on wells it has never seen.

The compounding of these two moves is the headline of the data-centric story. At the level of a single interval, augmentation lifted that two-hundred-thirty-six-patch band to **four thousand two hundred twelve patches**, raised the sinusoid-bearing patches from nineteen to **two thousand forty-six**, and grew the individual labelled sinusoids from thirty-two to **three thousand five hundred sixty-five** — a better-than-tenfold increase at the interval level, and a wholesale repair of the class balance. At the level of the full working set, the combination of overlapping windows and augmentation scaled the sinusoid corpus from roughly **nine hundred image-and-ground-truth pairs to over fifty-five thousand — a sixty-five-fold expansion**. Walk the same lineage in patch terms across the experiment progression and the earliest configuration's two hundred thirty-five patches became seventy-five thousand, a one-hundred-thirty-two-fold increase, as the patch geometry and stride were tuned alongside the augmentation factor.

The payoff lands exactly where the disease was. The degenerate model that posted one hundred percent classification error without augmentation dropped to **2.618 percent** with it. The auxiliary losses moved in lockstep — the Hungarian set-matching loss from 0.174 to 0.0135 and the parameter-regression loss from 0.575 to 0.062 — but the classification figure is the one that tells the story, because it is the difference between a model that never commits to the signal class and one that finds it reliably. No architectural change in this programme produced a swing of that magnitude. The data did.

Clean inputs first: why imputation is part of the data engineering

Before a patch is ever augmented, it has to be a faithful input, and borehole images arrive with holes. The microresistivity imaging tools read the borehole wall through a set of pads that do not cover the full circumference, leaving vertical null bands running the height of the image where no measurement exists. Those gaps cut straight through the sinusoids the detector is meant to trace, so whatever fills them becomes part of the curve the model learns. Imputation is therefore not a cosmetic pre-processing step in this domain — it is upstream model engineering, and it is unusually well-posed: the only question that matters is whether the fill keeps a sine wave a sine wave across the cut.



A high-resolution borehole image-log pad gap — the dead strip left between two different microresistivity imaging tools' pads — cuts a vertical null band through the unrolled borehole image, and whatever fills it becomes part of the sinusoid a detector traces — so the imputation question is well-posed: does the fill keep a sine wave a sine wave across the cut? Pick a method and the recovered fill redraws across the gap: KNN imputation (n_neighbors=5) interpolates along the curve and stays continuous (teal); the GAN inpaints locally-realistic texture that breaks the curve and exits at the wrong phase (the orange discontinuity is the argument); 1D-linear flattens it to a chord and leaves vertical-line artifacts; the iterative imputer stays continuous but is too slow for per-well runs. KNN won. The method ranking and compute markers (1D ~0.115 s vs KNN ~2.625 s on a 4 m interval; 1D ~11 s whole-well; KNN never finished a whole-well pass) are the article's own; the borehole image texture and the recovered-sinusoid curves are schematic.

The ladder above is the comparison we ran, and the verdict is counter-intuitive in a way worth dwelling on. A generative fill can produce locally-realistic borehole texture that looks convincing pixel-by-pixel and yet breaks the geometry — it hallucinates plausible rock where a continuous curve should be, and the sinusoid the detector traces snaps at the seam. A simple linear fill keeps things continuous but flattens the curve and leaves vertical-line artefacts. The method that won was a classical k-nearest-neighbours imputer with five neighbours: it preserves feature continuity across the null band, is cheap enough to run, and crucially does not invent geometry the instrument never measured. In a scarce-data regime the instinct is to reach for the most sophisticated model at every stage; here the

disciplined choice was the one that refused to fabricate. The augmentation regimen then operates on inputs that are already geometrically honest, which is why the ten variants per patch teach appearance robustness rather than amplifying a fabrication.

The real lesson: fourteen wells beat ninety-two thousand synthetic samples

Augmentation rescued the model from degeneracy. It did not, by itself, make the model good. The second half of the data-centric story is about which additions actually improve a scarce-data model — and it contains the single most important negative result of the engagement.

First, the positive result. Holding the architecture fixed and varying only the number of real wells in the training set traces a clean, decisive curve. With three wells, the model's classification error was **93.115 percent** — barely better than degenerate, and the Hungarian matching loss sat at 0.801. Six wells dropped the error to 18.370. Nine wells reached 1.055. Eleven wells, 0.817. And the full fourteen-well set landed the model at a **2.536 percent** classification error with a Hungarian loss of 0.015 — a more than fifty-fold reduction in matching loss from the three-well baseline. Each well of real, expert-labelled formation is worth more than any volume of manufactured variation, because each well carries genuinely new geological variation that augmentation cannot synthesise. The same pattern held on the regression side: moving from eight wells to eleven improved depth, dip, and azimuth error by roughly 0.007 in mean absolute error. Real diversity is the resource that actually moves the model.

Now the negative result, which is the one teams most need to internalise. Faced with a model that wanted more data and an operator that had no more wells to give, the tempting move is to manufacture volume — push the overlap and augmentation factors harder until the corpus is enormous. We tried it. One configuration inflated the training set to roughly **ninety-two thousand** overlapped-and-augmented samples. The model got **worse**. The extra volume was not extra information; it was the same handful of wells' worth of signal, copied and perturbed into a corpus large enough to look impressive and redundant enough to teach nothing new, while quietly biasing the model toward the over-represented intervals. The fix was not more synthetic volume but better data hygiene — relaxing the stride from forty to eighty to cut the redundancy, and validating on a genuinely held-out blind interval the model had never seen in any form. Scale, on its own, was a trap.

Why 25% is generous. The 14-task benchmark has four categories; a general LLM scores only on conceptual knowledge (5/10) and posts zero on data-grounded queries (0/4) and hallucination resistance (0/6) — the categories production cares about most. Slide the weighting from consumer fluency to production trust and Copilot's weighted score collapses (~34% → ~10%) while the domain-native WellBot holds ~94% → ~99%. The as-tested equal-points benchmark sits in between at 95.8% vs 25.0%. Per-category scores are the whitepaper's own; the weighting is an illustrative lens.

The instrument above makes the trade-off legible: how much you credit raw corpus size versus genuine, real-well diversity decides which strategy looks better, and any honest weighting that rewards out-of-distribution generalisation favours the fourteen real wells over the ninety-two-thousand-sample inflation. This is the canonical data-centric-AI lesson, stated in the sharpest possible terms by a real subsurface programme: **past a point, synthetic volume is not a substitute for real coverage, and treating it as one actively degrades the model.** Augmentation is a multiplier on the signal you have; it is not a generator of signal you lack. Confuse the two and you build a larger, slower, more confident, and less accurate detector.

The engineering that makes data-centric work reproducible

A data-centric programme only beats a model-centric one if its data decisions are auditable, because the entire argument rests on knowing precisely which corpus produced which number. That puts a hard requirement on the engineering underneath. Every generated dataset in this programme was a versioned, content-addressed artefact rather than a folder someone named by hand, so that "the fifty-five-thousand-pair set" referred to one specific frozen object and not to whichever variant happened to be on disk that week. The augmentation recipe — the six transforms, the ten-per-patch factor, the patch geometry, the stride — was itself part of the dataset's lineage, because changing any of those parameters produces a different corpus that will train a different model. A run referenced its dataset by hash; the hash referenced a frozen set of patches; the patches traced back through the recipe to the specific wells and depth intervals that produced them.

That lineage is what made the negative result trustworthy. When the ninety-two-thousand-sample corpus underperformed the leaner real-well set, we could prove it was the data and not a training fluke, because both runs were pinned to known datasets, known seeds, and self-describing checkpoints whose filenames encoded the learning rate, backbone, loss, and timestamp that produced them. The supervised configuration the search converged on — a from-scratch ResNet-10 backbone, a four-layer transformer encoder and decoder, AdamW at a tuned learning rate, a combined focal-and-L1 loss weighting the classification term above the regression term, and early stopping — was a tracked, reproducible record rather than folklore. In a data-centric programme the experiment-tracking discipline is not optional infrastructure bolted on at the end; it is the instrument that lets you distinguish a real data gain from a lucky one, and it is what allowed the operator's own engineers to inherit the pipeline and retrain on new wells without rediscovering every decision from scratch.

Why this generalises beyond boreholes

Strip away the borehole-imaging specifics and the structure of this problem recurs across every proprietary domain where labelled data is scarce and expensive: clinical imaging from a single hospital network, defect detection on a niche manufacturing line, any computer-vision task where the corpus is a few hundred expert-labelled examples and no public dataset matches the distribution. The instinct in those settings is to treat scarcity as a

modelling problem and reach for a bigger architecture or a foundation model fine-tune. The evidence from this engagement points the other way. The model was a Detection-Transformer variant with a deliberately small backbone — a ResNet-10 trained from scratch, chosen specifically because a larger backbone overfit the scarce data — and the architecture search, while necessary, produced gains an order of magnitude smaller than the data work did.

What moved the number, in order of impact, was: cleaning the inputs so the geometry was honest, manufacturing more legitimate training examples through overlapping windows and physically-motivated augmentation, repairing a brutal class imbalance by augmenting only the signal class, and — above all — adding real wells while refusing to mistake synthetic volume for them. That is a data-engineering programme, not a modelling one. In scarce proprietary domains the durable advantage is not the model you can download; it is the discipline with which you turn a small body of proprietary, expert-labelled data into a training distribution that teaches. The operators we have worked with across the Middle East and the United States who succeed with applied AI are the ones who treat their scarce labelled data as the asset and the model as the multiplier — never the reverse.

What good looks like

For a chief data officer, an ML platform lead, or a head of data science evaluating an AI programme in a scarce proprietary domain, the questions that matter are not about the architecture. They are about whether the data engineering was done as engineering:

- Is the input conditioning — imputation, normalisation, gap-filling — chosen to preserve the geometry the model has to learn, rather than to produce convincing-looking pixels?
- Is augmentation modelling the domain's real corruption manifold and rebalancing the signal class, rather than adding generic noise to the whole set?
- Is the patching and stride schedule treated as a tuned hyperparameter, with redundancy actively managed rather than maximised?
- Is the team measuring the marginal value of a real, expert-labelled example against the marginal value of a synthetic one — and acting on the answer when synthetic volume stops helping?
- Is every model evaluated on a genuinely held-out, out-of-distribution interval, so that corpus inflation cannot disguise itself as improvement?

If the answers are yes, the programme is winning where scarce-data programmes are won. If the instinct is still to reach for a bigger model and a larger synthetic corpus, the programme is optimising the cheap half of the problem and starving the half that decides the outcome.

WHAT THIS WHITEPAPER ARGUES

1. In scarce proprietary domains the binding constraint is data, not architecture — model engineering becomes data engineering, and the decisive gains are won data-side.
2. Overlapping-window patching plus a six-transform, ten-per-patch augmentation regimen scaled the sinusoid set ~65x (≈900 → 55,000 pairs; one interval's 236 patches → 4,212) and rebalanced a brutal minority class.
3. Augmentation collapsed a degenerate classification error from 100% to 2.618% — a swing no architecture change in the programme came close to.
4. Adding real wells (3 → 14) drove classification error 93.115% → 2.536%; each well of expert-labelled formation carries variation augmentation cannot synthesise.
5. Inflating the corpus to ~92,000 synthetic samples made the model WORSE — past a point, synthetic volume is redundancy, not information; real coverage and out-of-distribution validation are what generalise.

References

EarthScan, 2023 Detection-Transformer-based borehole-feature picking under well scarcity — formation-evaluation programme, mid-sized Middle East carbonate operator. Internal technical report, EarthScan.

Carion et al., 2020 N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko. End-to-End Object Detection with Transformers (DETR). ECCV 2020. Architectural basis for the set-prediction detection approach. https://arxiv.org/abs/2005.12872

Hendrycks et al., 2020 D. Hendrycks, N. Mu, E. Wilson, et al. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. ICLR 2020. The consistency-augmentation principle underlying robust image-degradation training. https://arxiv.org/abs/1912.02781

Northcutt et al., 2021 C. Northcutt, A. Athalye, J. Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. NeurIPS 2021. The data-centric argument that label and data quality, not model size, dominate real-world performance. https://arxiv.org/abs/2103.14749

EARTHSCAN

Data-Centric AI Under Extreme Well Scarcity: How 235 Patches Became 75,000 and Why 14 Wells Beat Synthetic Overload

Published March 2025. Generated 2026-07-22.
https://www.earthscan.io/whitepapers/data-centric-ai-well-scarcity-augmentation-65x