Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Does a Brain-Tumour Segmentation Model Notice When a Weak Telemedicine Link Has Degraded Its Input?

Adrija Das

adrija.das.cs@gmail.com

University of Utah

Manabendra Maiti

Techno International New Town

Research Article

Keywords: Calibration, failure detection, image compression, packet loss, segmentation uncertainty, teleradiology

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11803554/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

Does a Brain-Tumour Segmentation Model Notice When a Weak Telemedicine Link Has Degraded Its Input?

Adrija Das
Independent Researcher
formerly Kahlert School of Computing, University of Utah
Salt Lake City, UT, USA
adrija.das.cs@gmail.com (corresponding author)

Dr. Manabendra Maiti
Professor, *Dept. of Electronics and Communication Engineering*
Techno International New Town, Block DG 1/1, Action Area I,
New Town, Kolkata, West Bengal 700156, India
dr.manabendra.maiti@tict.edu.in

***Abstract*—A clinic without a specialist may send scans over a slow or lossy link to a hub where AI segments them. We asked whether the model's own confidence warns when the transfer has damaged the scan. A simulator, validated on phantoms, turns a link and a deadline into what arrives: rate-controlled compression over reliable transport (assumed tiers and LTE traces), messaging-app recompression, and, as a worst case, real-time transport that loses image strips. We trained five U-Nets (one used alone, all five as an ensemble), a Monte Carlo dropout U-Net and nnU-Net on 70 patients of a public glioma MRI dataset, and tested them on 30 held-out patients (1320 transmissions per model). Over reliable transport, mean Dice per condition stayed within 0.027 of clean; for studies sent below 0.3 bits per pixel (down to 0.11), it changed by -0.011 to -0.030. Losing strips was harmful. Row interpolation removed 92–99% of the harm under independent loss and 59–83% under bursts. On failures (Dice lost ≥ 0.10), confidence in the predicted tumour barely fell, and after temperature scaling every model was still overconfident (ECE 0.099–0.176). Disagreement between test-time-augmentation passes ranked U-Net failures better than U-Net's confidence did (area under the risk–coverage curve 0.109 against 0.141). Ensemble and nnU-Net confidence did as well as uncertainty; a single ensemble member's did not. Receiver-side image quality detected transmission damage (AUROC 0.91 for U-Net) but missed patients who were hard before transfer. A combined score did well on both.**

***Index Terms*—Calibration, failure detection, image compression, packet loss, segmentation uncertainty, teleradiology.**

I. Introduction

Teleradiology lets a clinic without an on-site specialist send images for remote reading [1]. In rural areas the connection is often a cellular link whose capacity varies. To meet a deadline, the sender may compress or downscale a study, a messaging app may recompress it [2], and a transfer without retransmission may lose parts of it. A human reader may notice the damage. An AI model may not: deep networks are sensitive to corruptions they were not trained on [3], [4], [5], and their uncertainty becomes less reliable under shift [6].

We place the AI at the receiving hub, after transfer, because that is where transmission damage matters. The hub also knows how each study travelled: codec, quality factor, bits per pixel and lost packets. A model running at the clinic before upload would never see the damage.

Earlier studies apply transmission impairments or clinical degradations directly to images [7], [8]. A clinic experiences a *link*, and the damage depends on how the sender adapts to it. We connect the two. For four models and four uncertainty methods, we ask how accuracy and calibration change with the link and the transport, whether a model's confidence falls when its output fails, and which per-study signal best tells the hub to re-check a result. We measure the last with the failure-detection metrics now standard for segmentation [9], [10]. Our contributions are a phantom-validated simulator that turns a link into received images; an evaluation of four models on 30 patients unused in training or model selection; and a risk–coverage comparison of confidence, uncertainty, transfer metadata and a receiver-side image-quality score as warnings.

II. Related work

Corruption benchmarks sweep generic perturbations [4], [5], [11]. We derive ours from a link and from the sender's rate control. Lossy JPEG 2000 compression up to about $20\times$ did not harm 3-D segmentation [12], and moderate compression did not harm mammogram classification [13]. Downsizing and compression have been studied for classifiers serving remote clinics [14], and chest X-rays sent through a messaging app made predictions unstable [2]. Thanoun *et al.* simulated JPEG coding, channel noise and bursty loss for kidney-tumour segmentation and found loss far more damaging than compression [7]. They modelled no sender adaptation, retransmission, concealment or confidence. FrugalVision adapts a learned codec to measured 5G traces but judges the result by PSNR rather than by a model's output [15]. MC-dropout uncertainty flagged failures on degraded synthetic brain-tumour data [8].

For failure detection, Jaeger *et al.* argue for risk–coverage curves and their area (AURC) as a unified metric [9]. In segmentation, Zenk *et al.* showed that the way pixel confidence is aggregated into a per-image score matters, and that pairwise Dice between ensemble members is a strong simple baseline [10]. ValUES separates prediction model, uncertainty measure and aggregation, and finds test-time augmentation (TTA) a light-weight option [16]. QU-BraTS scores brain-tumour uncertainty [17], and learned quality-control networks predict

Dice for brain-tumour segmentations directly [18]. Under synthetic acquisition shift, a single nnU-Net's confidence stayed nearly flat while disagreement between ensemble members rose [19]. We use TTA [20], deep ensembles [21], MC dropout [22] and nnU-Net [23]. We measure calibration by ECE [24], [25] and by the average calibration error (ACE), which weights bins equally [26]. Conformal risk control bounds an expected loss [27] and has been applied to lesion segmentation [28].

On the network side, DICOM transfer runs over TCP [29], with JPEG 2000 and HTJ2K among its transfer syntaxes [30]. Loss-based throughput models [31] describe Reno but not BBR [32]. Recorded cellular traces drive emulators such as Mahimahi [33]. Replay fidelity [34] and cellular throughput [35] are still being studied.

III. Methods

A. Data, split and models

We used the public lower-grade glioma dataset of Buda *et al.* [36], [37], [38]: 110 patients from TCGA-LGG [39], [40], with 2-D axial slices of 256×256 pixels in three 8-bit channels (pre-contrast, FLAIR, post-contrast) and FLAIR-abnormality masks. We fixed a split stratified by acquisition site: 70 training, 10 validation and 30 test patients. The test patients were used only for the final evaluation. They come from the same sites as the training patients, so the test is internal, not external.

We trained five U-Nets with the released architecture [41], [42] from scratch, each with a different seed, using the released recipe with Dice loss (Section S1). We kept the epoch with the best validation Dice (0.814 for the strongest seed). Seed 0 is the single U-Net; all five form the deep ensemble. For MC dropout, a sixth U-Net adds 2-D dropout ($p = 0.2$) after the two deepest encoder blocks, the bottleneck and the two deepest decoder blocks (best validation Dice 0.828). nnU-Net v2 [23], in its 2-D configuration, was trained on the same training and validation patients for at most 250 epochs, keeping the best checkpoint. Each model uses its own preprocessing. Dice always comes from a model's deterministic own-confidence map (Section III-C), never from a TTA mean: we map it back to the original grid, threshold it at 0.5 and score it with 3-D Dice, without post-processing.

B. Link and transport

A study is sent as one greyscale image per slice and sequence and must arrive within a deadline. Our default is 60 s, about 2% of the 46-minute mean report turnaround reported for intensive-care teleradiology [43]; we also test 15 s and 300 s for tighter and slower settings. The median test study has 5.9 million 8-bit pixels (5.9 MB raw), and lossless coding needed a median of 3.07 bits per pixel. We state budgets in bits per pixel (bpp; raw = 8) because that unit does not depend on study size. Over 60 s, 1 bpp is 98 kb/s for the median study. A native study with 512×512 images and twice the slices has eight times the pixels and needs eight times the throughput for the same bpp, and its 12–16-bit data may need more bpp for the same accuracy, which we did not test.

The sender picks one setting per study: lossless if the study fits the byte budget, otherwise the resolution scale $\{1, 0.75, 0.5, 0.375, 0.25\}$ and JPEG quality $\{95, \ldots, 5\}$ that maximize mean PSNR within it. Over reliable transport, the budget has two sources. Four assumed tiers (capacity 10, 2, 0.5 and 0.1 Mb/s; round-trip time 50–600 ms; loss 0.1–8%) set it through the Mathis model [31], giving 9012, 950, 274 and 84 kb/s. Modern stacks are not loss-based [32], so we treat these tiers as illustrations. Recorded LTE uplink traces captured while driving (AT&T and Verizon, recorded for Sprout over a decade ago [44] and distributed with Mahimahi [33]) give the second: the bytes an ideal capacity-limited sender can deliver in a window of the trace starting at a random time, two windows per patient. Driving traces likely overstate outages and capacity swings for a clinic at a fixed site, so this arm is pessimistic. For a deployer, we also compute the codec choice for 16–1024 kb/s over 60 s, which runs from 0.16 bpp for the median study (0.09 for the largest) to above lossless. The messaging app recompresses the study as an RGB JPEG (quality 70, 4:2:0), once or twice (forwarded).

Real-time transport is our worst case. Without retransmission, a lost packet removes one 8-row strip of one image. We give each strip a restart marker, which is generous to the image. Strips are lost independently or in bursts (Gilbert–Elliott [45], [46], mean burst 4) and are zero-filled or concealed by linear interpolation between the nearest received rows. We test both fills under both kinds of loss. Each tier's loss rate is also applied without retransmission. At T2 this is the 1% strip loss of the loss sweep with the same draws, so we count it once. We keep this arm as a worst case because store-and-forward teleradiology retransmits [29]. A JPEG quality sweep at full resolution completes the conditions. Before using any patient data, we checked every component on synthetic phantoms with known answers. 35 of 35 checks passed, including two negative controls that must fail and did (Table S7).

C. Confidence, uncertainty and calibration

Each model gives a deterministic *own-confidence* map (a single pass, the member mean for the ensemble, dropout off for MC dropout) and an uncertainty estimate from several passes: 8-pass TTA (flip $\times$ four integer shifts) for U-Net and nnU-Net, the 5 members, and 10 MC-dropout passes. For U-Net we add a *degradation-aware* TTA with 7 re-degraded copies of the received study (JPEG 75/50/30; strip loss 2% and 5%, two draws each). Per-voxel uncertainty H is the binary entropy of the mean probability. Following [10], [16], we aggregate per study: mean probability inside the predicted tumour (*confidence*); pairwise Dice between passes; the uncertainty ratio $U_\tau = |\{H > \tau\}|/(|\{H > \tau\}| + |\{\hat{y} = 1\}|)$ over voxels, with $\tau \in \{0.3, 0.5, 0.7\}$ bit and $\hat{y}$ the predicted mask; mean entropy inside the predicted tumour [47] and in a ± 2 px band around its boundary; the predicted-volume change against the mean prediction; and the volume coefficient of variation across passes. We also report the QU-BraTS score [17]. Calibration uses the predicted-class confidence in 15 bins (ECE and ACE [26]), in the union of prediction and ground

truth dilated by 5 px; the supplement also reports it over the whole image. We fitted one temperature per model on the clean validation patients only, minimizing the negative log-likelihood over $T \in [0.05, 200]$ on logits: the pre-sigmoid output for the U-Nets, the difference of the two class logits for nnU-Net, and the log-odds of the member mean for the ensemble.

D. Failure detection and statistics

Each test patient was sent under 31 named conditions, clean and 30 degraded (Table S4). We ran conditions with random draws (independent, bursty and concealed strip loss, real-time tiers, trace windows) twice, with different seeds or windows. That adds 14 transmissions, for 44 degraded transmissions per patient and 1320 per model. A study that missed its deadline (6, all after trace outages) did not reach the hub in time. We report it as *late*, a delivery failure, and leave it out of the accuracy and failure analyses. Every other received study is a case (1314 per model). We use two views of risk. The *absolute* risk, $1-$Dice of the received result, is primary, as in [10]. The *transmission* risk is the Dice lost relative to the same patient's clean result. The warnings are the aggregations above, the lost-strip fraction, and three quality regressions (ridge, standardized features, fitted leave-one-patient-out on the same target). The *metadata* regression uses lost fraction, log bits per pixel, scale, quality, lossless and messaging. The *receiver image-quality* regression uses only the received FLAIR: block-edge ratio at 8 and 16 px [48], sharpness and fraction of black rows. The *combined* regression uses both sets plus confidence, pairwise Dice and $U_{0.5}$. AURC [9] is the primary metric. AUROC for a relative failure (Dice lost ≥ 0.10, and 0.05–0.20) and for an absolute failure (Dice < 0.7) are secondary. Intervals come from a patient-cluster bootstrap, and paired comparisons use pairwise Dice as the uncertainty score. For accuracy, the patient is the unit and seeds are averaged. We report BCa intervals [49], two one-sided tests (TOST) for equivalence within ± 0.01 Dice, and Wilcoxon tests [50], each Holm-corrected [51].

IV. Results

A. Accuracy and transport

On clean test scans, mean Dice was 0.817 (95% CI 0.720–0.864) for U-Net, 0.837 for the ensemble, 0.830 for MC dropout and 0.865 (0.815–0.893) for nnU-Net. Hard patients pull the means down: 4 scored below 0.7 with U-Net, and one scored below 0.5 with every model. Medians were 0.880–0.910.

Reliable transport barely changed accuracy (Table I, Fig. 1). Dice was equivalent to clean in 9 of 13 reliable-transport conditions for U-Net, 10 for the ensemble, 10 for MC dropout and 8 for nnU-Net. Where equivalence was not shown, the changes were small: at most 0.018 for the U-Net models and 0.027 for nnU-Net (T4 with a 15 s deadline). Below 0.3 bpp (26 studies, down to 0.11 bpp, about 70:1), mean Dice changed by -0.011 (U-Net) to -0.030 (nnU-Net). No U-Net, ensemble or MC-dropout study that arrived in time failed; 2 nnU-Net studies did. Rate control kept full resolution for at least 95% of studies down to 0.4 bpp (about 20:1, or 39 kb/s for the median study). Below that it lowered resolution before quality: only 1 of 360 settings went below JPEG quality 20. On the LTE traces, 6 of 180 windows delivered nothing before the deadline, and the floor setting arrived late. Counted as Dice failures, these outages would be all 6 trace failures (Table S9). No window that delivered in time caused a failure. The messaging app cost less than one Dice point, sent once or forwarded, but equivalence within ± 0.01 was not shown.

Without retransmission, loss hurt. At T4 (8% packet loss), Dice changed by -0.108 for U-Net and -0.135 for nnU-Net. Bursty loss at 5% did more harm than independent loss at the same rate. Concealing lost strips by row interpolation removed 92–99% of the harm under independent loss but only 59–83% under bursty loss, where neighbouring strips are lost together and interpolation must bridge wider gaps. The other harmful condition was heavy compression at full resolution (JPEG quality 5: -0.153 for U-Net, -0.172 for nnU-Net; Table S4).

B. Confidence on failures

In absolute terms, confidence barely moved. Failures (Dice lost ≥ 0.10) were 111 of 1314 degraded studies for U-Net, from 22 patients (other models: Table II), and no patient contributed more than 18. On these failures Dice changed by -0.212, -0.200, -0.220 and -0.224, while mean confidence in the predicted tumour changed by only -0.005, -0.064, -0.011 and -0.021 (from 0.982, 0.934, 0.984 and 0.992 on clean scans). The ensemble mean, the least confident map, moved most. Temperature scaling on the logits of clean validation scans ($T = 6.34$, 3.29, 9.12 and 10.03, none at the search bound) reduced ECE near the tumour on clean test scans for the U-Net models (U-Net: 0.168 to 0.047). It left nnU-Net under-confident there (confidence minus accuracy -0.138). On failures, every scaled model was still overconfident (ECE 0.176, 0.135, 0.162 and 0.099; Fig. S1). No single temperature fits clean and damaged scans.

C. Comparison of warnings

Table II and Fig. S4 compare the warnings. For absolute risk, pairwise Dice between passes reached AURC 0.109 (0.091–0.142) for U-Net, against 0.141 (0.108–0.191) for its confidence (paired difference 0.032, 0.012–0.061). That is close to the oracle (0.098) and far from random (0.211). Uncertainty also beat confidence for MC dropout (difference 0.026, 0.012–0.043). For the ensemble and nnU-Net, confidence ranked failures as well as uncertainty did (differences -0.004 and 0.004). The ensemble's confidence map, however, is the mean of five members and already contains their disagreement. Each member's own confidence ranked failures like the single U-Net's (AURC 0.127–0.154, AUROC 0.73–0.78), against 0.091 and 0.87 for the mean (Table S8). The choice of τ mattered only for MC dropout (AURC 0.095 at 0.3 bit, 0.160 at 0.7 bit), and failure thresholds of 0.05–0.20 gave the same picture (Section S3).

TABLE I

CHANGE IN MEAN DICE FROM CLEAN (×100), BY CONDITION AND MODEL; FAILURES (DICE LOST ≥ 0.10) IN BRACKETS, AND STUDIES THAT MISSED THE DEADLINE ("LATE", EXCLUDED FROM THE MEAN). †EQUIVALENT TO CLEAN WITHIN ±0.01 (TOST, HOLM $p < 0.05$); *DECREASE (WILCOXON, HOLM $p < 0.05$). FULL TABLE IN THE SUPPLEMENT.

Condition	U-Net	Ensemble	MC dropout	nnU-Net
Clean Dice, mean	0.817	0.837	0.830	0.865
Reliable transport				
T1, T2, 60 s	0.0†	0.0†	0.0†	0.0†
T3, 60 s	−0.1†	0.0†	+0.1†	−0.1†
T4, 60 s	+0.2	+0.1†	0.0†	−0.4 (1)
T4, 15 s	−1.0	−0.7	−1.8*	−2.7* (1)
T4, 300 s	−0.2†	−0.2†	0.0†	0.0†
LTE trace A, 60 s	0.0† (1 late)	0.0† (1 late)	0.0† (1 late)	0.0† (1 late)
LTE trace A, 15 s	−0.2† (5 late)	−0.1† (5 late)	+0.1† (5 late)	−0.1† (5 late)
LTE trace B, 15 s	0.0†	0.0†	0.0†	0.0†
Messaging app, sent once	−0.5	−0.4	−0.4	−0.4
Messaging app, forwarded	−0.5	−0.5	−0.3	−0.4
Real-time transport, worst case (no retransmission)				
T2 (1% loss)	−1.5* (1)	−1.0*	−1.7* (2)	−1.5* (2)
T3 (3% loss)	−3.1* (6)	−2.4* (2)	−4.5* (6)	−3.6* (6)
T4 (8% loss)	−10.8* (22)	−9.8* (23)	−13.6* (29)	−13.5* (23)
5% strips, zero-filled	−7.1* (14)	−6.2* (12)	−9.4* (20)	−7.6* (12)
5% strips, concealed	−0.2	−0.2	−0.3*	−0.6*
5% strips, bursty	−8.2* (14)	−8.6* (19)	−11.8* (28)	−11.0* (24)
5% strips, bursty, concealed	−2.4* (3)	−2.7* (3)	−2.0* (2)	−4.6* (6)
10% strips, zero-filled	−15.5* (34)	−14.0* (28)	−17.9* (39)	−17.1* (34)
10% strips, concealed	−0.2	−0.3	−0.5*	−1.3* (1)

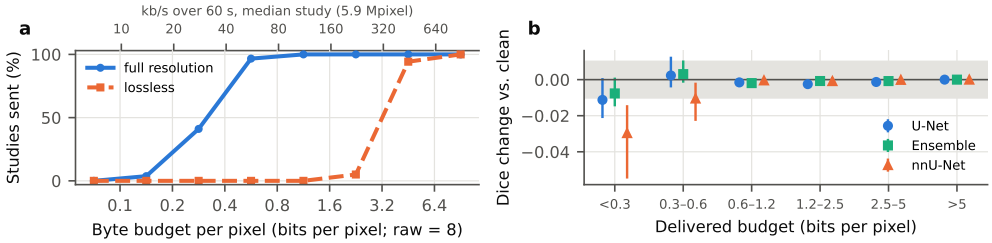


Fig. 1. (a) Rate-control choice (codec only) against the byte budget in bits per pixel for a 60 s deadline, across the 30 test studies; top axis: throughput for the median study. (b) Measured Dice change from clean against the delivered budget for all reliable-transport studies delivered in time (mean, 95% bootstrap interval; band: ±0.01).

TABLE II

FAILURE DETECTION OVER ALL DEGRADED STUDIES: AURC FOR ABSOLUTE RISK (1−DICE; LOWER IS BETTER; BEST PER MODEL IN BOLD) WITH AUROC FOR DICE LOST ≥ 0.10 IN BRACKETS. REGRESSIONS ARE FITTED LEAVE-ONE-PATIENT-OUT. INTERVALS FOR EVERY ENTRY ARE IN SUPPLEMENTARY TABLE S3.

Warning score	U-Net	Ensemble	MC dropout	nnU-Net
Mean confidence in predicted tumour	0.141 (0.77)	**0.091** (0.87)	0.120 (0.71)	0.097 (0.87)
Pairwise Dice disagreement	**0.109** (0.86)	0.095 (0.89)	**0.094** (0.85)	**0.093** (0.90)
Uncertainty ratio, $H > 0.5$ bit	**0.109** (0.87)	0.096 (0.89)	0.125 (0.85)	**0.093** (0.90)
Volume change vs. mean prediction	0.132 (0.72)	n/a	0.194 (0.79)	0.130 (0.78)
Metadata regression	0.214 (0.85)	0.188 (0.87)	0.198 (0.86)	0.170 (0.79)
Receiver image-quality regression	0.250 (0.91)	0.234 (0.91)	0.213 (0.92)	0.177 (0.88)
Combined regression	0.111 (0.92)	0.094 (0.91)	0.116 (0.91)	**0.093** (0.94)
Oracle / random	0.098 / 0.211	0.087 / 0.188	0.088 / 0.205	0.086 / 0.170
Failures: studies / patients	111 / 22	98 / 22	148 / 24	137 / 25

For absolute risk, metadata and receiver image quality were no better than random for U-Net (AURC 0.214 and 0.250 against 0.211). They cannot see patients who are hard before any transfer, such as the 4 with clean U-Net Dice < 0.7. Without the two hardest patients (clean U-Net Dice 0.06 and 0.46; 88 of U-Net's 251 absolute failures), they stayed close to random (U-Net: 0.165 and 0.151 against 0.173) and were the weakest warnings for every model (Table S10). For transmission risk the link-side signals caught up. For U-Net's 111 failures, the receiver image-quality regression reached AUROC 0.91 (0.89–0.93) and metadata 0.85 (0.82–0.88), against 0.86 (0.78–0.93) for pairwise Dice and 0.77 (0.69–0.84) for confidence. The combined regression was close to the best warning on both views (U-Net: AURC 0.111, AUROC 0.92, 0.85–0.96). To see whether the regressions transfer to degradations they were not fitted on, we refitted them on JPEG,

strip-loss, tier and deadline conditions only and tested them on the messaging app, concealment, bursty loss and traces. The U-Net image-quality regression kept its AUROC (0.89; 0.88 when fitted on all families). Metadata fell to 0.60 (0.69; Table S11).

Uncertainty concentrated at the tumour boundary: for U-Net, high-entropy voxels were $31\times$ over-represented within 2 px of it. It barely marked lost strips ($1.3\times$). Degradation-aware TTA did not improve detection (AURC 0.112) and took 19 s per study, against 2.9 s for 8-pass TTA and 4.9 s for nnU-Net with TTA (one T4 GPU).

V. Discussion

How the scan travelled mattered more than how slow the link was. With retransmission and rate control, the models tolerated even the tightest setting, T4 with a 15 s deadline, where 84 kb/s for 15 s (21 kb/s averaged over a minute) left about 0.1 bits per pixel for the largest studies; the LTE traces behaved alike. The damage came from losing data, through real-time transport or through an outage that stops delivery before the deadline. A gateway should therefore keep retransmission, conceal the losses it cannot avoid (less effective for bursts), and report a missed deadline rather than silently forward a floor-quality study. In [7], too, loss harmed segmentation more than compression, and far more than here: 1% loss of scattered 4×4-pixel blocks roughly halved Dice on kidney CT, against -0.015 for our 1% strip loss. Loss pattern, modality and model all differ, though.

In absolute terms, every model stayed confident on failures, and temperature scaling on clean data did not change that, so a reader cannot take a high confidence value as reassurance. As a ranking, nnU-Net's confidence was informative, and so was the ensemble's, though not its members' (Table S8). We averaged confidence inside the predicted tumour because that ranks failures better than an image-wide average [10], [52]. Even so, the single U-Net and MC dropout needed pass-to-pass disagreement. Both were trained with Dice loss and batch normalization, which make U-Nets overconfident, while ensembling improves calibration [47]. Our AUROCs for a relative failure with pairwise Dice, 0.85 (MC dropout) to 0.90 (nnU-Net), are in line with the 0.843 that MC-dropout entropy from an Attention U-Net reached under synthetic Gaussian noise for slices with whole-tumour Dice below 0.65 [8].

Link-side signals and model uncertainty catch different failures. Metadata and image quality are cheap and detect what the link did, but not a patient who is hard for the model, just as generic no-reference quality metrics barely predicted segmentation failure on retinal OCT [53]. Model uncertainty detects what the model cannot do, including on intact scans. A hub needs both. A combined regression is a simple way to get them, and combining scores also helped in a large classification benchmark [54]. We propose three displays, not yet tested with readers: lost strips marked on the received scan, an uncertainty-only overlay, and a per-study "check again" flag from the combined score, with its threshold set by conformal risk control [27], [28] rather than by tuning. Set on half the test patients and tested on the other half (1000 splits), a split-conformal threshold missed 10% of U-Net failures for a 10% target, with a 5th–95th percentile range of 2–21% across splits, and flagged 22% of studies (Table S12). Whether these displays help clinicians, and whether they reduce automation bias [55], [56], needs a reader study.

The study has several limits. The images are the dataset's 8-bit derivatives of 2-D slices, not native 12–16-bit DICOM, and we did not test JPEG 2000 or HTJ2K, so the bits per pixel at which compression hurts may differ. We used one dataset and one task. Trace budgets assume an ideal sender and come from US LTE traces recorded while driving, not from clinic links. The tiers are assumptions, and we did not emulate packets. The U-Nets were trained with Dice loss only, and the MC-dropout model was still improving at its last epochs. The 30 test patients come from one split that shares acquisition sites with training. No clinician evaluated the displays.

VI. Conclusion

Over reliable transport with rate control, every model's mean Dice stayed within 0.027 of clean in every condition, and changed by -0.011 to -0.030 for studies sent below 0.3 bits per pixel. Losing data harmed accuracy, and concealment recovered most of that harm, less so under bursty loss. Confidence stayed high on failures, even after temperature scaling. Pass-to-pass disagreement, or uncertainty combined with link-side signals, ranked failures best.

Data: TCGA-LGG data were obtained through TCIA [40], [39], de-identified, and used under its data-use terms; we used the public derivative of [36], [38]. LTE traces were recorded by Winstein et al. [44] and are distributed with Mahimahi [33].

References

[1] G. Humphreys, "Leveraging teleradiology with artificial intelligence," *Bull. World Health Organ.*, vol. 103, no. 2, pp. 86–87, 2025.

[2] M. Antony, S. T. Kakileti, R. Shah, S. Sahoo, C. Bhattacharyya, and G. Manjunath, "Challenges of AI driven diagnosis of chest X-rays transmitted through smart phones: A case study in COVID-19," *Sci. Rep.*, vol. 13, art. 18102, 2023.

[3] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *Proc. QoMEX*, 2016, pp. 1–6.

[4] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proc. ICLR*, 2019.

[5] L. Boone *et al.*, "ROOD-MRI: Benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in MRI," *NeuroImage*, vol. 278, art. 120289, 2023.

[6] Y. Ovadia *et al.*, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Proc. NeurIPS*, 2019, pp. 13991–14002.

[7] M. Y. Thanoun, M. J. M. Zedan, A. M. A. Sabaawi, A. A. Mohammed, E. Elbasi, and M. A. Zulkifley, "HUM-Net: A robust semantic segmentation model for medical imaging under wireless channel degradations," *J. Européen des Systèmes Automatisés*, vol. 59, no. 8, pp. 2179–2195, 2026.

[8] P. Kaliaperumal and M. Kaliaperumal, "Trustworthy medical segmentation: Uncertainty-aware U-Net evaluation under clinical image degradation," arXiv:2607.22727, 2026.

[9] P. F. Jaeger, C. T. Lüth, L. Klein, and T. J. Bungert, "A call to reflect on evaluation practices for failure detection in image classification," in *Proc. ICLR*, 2023.

[10] M. Zenk *et al.*, "Comparative benchmarking of failure detection methods in medical image segmentation: Unveiling the role of confidence aggregation," *Med. Image Anal.*, vol. 101, art. 103392, 2025.

[11] F. Di Salvo, S. Doerrich, and C. Ledig, "MedMNIST-C: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions," in *MICCAI Workshop ADSMI*, 2024, arXiv:2406.17536.

[12] A. Kurmukov, B. Zavolovich, A. Dalechina, V. Proskurov, and B. Shirokikh, "The effect of lossy compression on 3D medical images segmentation with deep learning," arXiv:2409.16733, 2024.

[13] Y.-Y. Jo *et al.*, "Impact of image compression on deep learning-based mammogram classification," *Sci. Rep.*, vol. 11, art. 7924, 2021.

[14] E. M. Hung, R. C. da Silva, A. O. O. dos Reis, and D. Akogo, "Image downsizing and compression impacts on AI-based medical image classification," *J. Commun. Inf. Syst.*, vol. 41, no. 1, pp. 16–25, 2026.

[15] A. Guidoum and A. E. Lafkir, "FrugalVision: Adaptive medical image compression for resource-constrained IoT environments," *Control Eng. Appl. Inform.*, vol. 28, no. 2, pp. 75–86, 2026.

[16] K.-C. Kahl, C. T. Lüth, M. Zenk, K. Maier-Hein, and P. F. Jaeger, "ValUES: A framework for systematic validation of uncertainty estimation in semantic segmentation," in *Proc. ICLR*, 2024.

[17] R. Mehta *et al.*, "QU-BraTS: MICCAI BraTS 2020 challenge on quantifying uncertainty in brain tumor segmentation—Analysis of ranking scores and benchmarking results," *J. Mach. Learn. Biomed. Imaging*, vol. 2022, art. 026, 2022.

[18] P. Qiu, S. Chakrabarty, P. Nguyen, S. S. Ghosh, and A. Sotiras, "QCResUNet: Joint subject-level and voxel-level segmentation quality prediction," *Med. Image Anal.*, vol. 107, art. 103718, 2025.

[19] R. D. Shet, C. Liang, and L. Zhang, "Reliability analysis for BraTS-GoAT segmentation: A controlled robustness study of deep-ensemble uncertainty," arXiv:2608.13223, 2026.

[20] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, 2019.

[21] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. NeurIPS*, 2017, pp. 6402–6413.

[22] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. ICML*, 2016, pp. 1050–1059.

[23] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[24] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in *Proc. AAAI*, 2015, pp. 2901–2907.

[25] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017, pp. 1321–1330.

[26] T. Barfoot, L. C. Garcia-Peraza-Herrera, B. Glocker, and T. Vercauteren, "Average calibration error: A differentiable loss for improved reliability in image segmentation," in *Proc. MICCAI*, 2024, doi:10.1007/978-3-031-72114-4_14.

[27] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," in *Proc. ICLR*, 2024.

[28] B. Tan *et al.*, "Conformal lesion segmentation for 3D medical images," arXiv:2510.17897, 2025.

[29] NEMA, *Digital Imaging and Communications in Medicine (DICOM) PS3.8: Network Communication Support for Message Exchange*, 2026c.

[30] NEMA, *DICOM PS3.5: Data Structures and Encoding*, Sect. 10.18, Transfer Syntaxes for High-Throughput JPEG 2000 Compression, 2026c.

[31] M. Mathis, J. Semke, J. Mahdavi, and T. Ott, "The macroscopic behavior of the TCP congestion avoidance algorithm," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 27, no. 3, pp. 67–82, 1997.

[32] N. Cardwell, Y. Cheng, C. S. Gunn, S. H. Yeganeh, and V. Jacobson, "BBR: Congestion-based congestion control," *ACM Queue*, vol. 14, no. 5, pp. 20–53, 2016.

[33] R. Netravali *et al.*, "Mahimahi: Accurate record-and-replay for HTTP," in *Proc. USENIX ATC*, 2015, pp. 417–429.

[34] W. Sentosa, B. Chandrasekaran, P. B. Godfrey, and H. Hassanieh, "CellReplay: Towards accurate record-and-replay for cellular networks," in *Proc. USENIX NSDI*, 2025, pp. 1169–1186.

[35] A. Narayanan *et al.*, "Lumos5G: Mapping and predicting commercial mmWave 5G throughput," in *Proc. ACM IMC*, 2020, pp. 176–193.

[36] M. Buda, A. Saha, and M. A. Mazurowski, "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm," *Comput. Biol. Med.*, vol. 109, pp. 218–225, 2019.

[37] M. A. Mazurowski, K. Clark, N. M. Czarnek, P. Shamsesfandabadi, K. B. Peters, and A. Saha, "Radiogenomics of lower-grade glioma: Algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with The Cancer Genome Atlas data," *J. Neuro-Oncol.*, vol. 133, no. 1, pp. 27–35, 2017.

[38] M. Buda, "LGG segmentation dataset (Brain MRI segmentation)," Kaggle. https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation

[39] K. Clark *et al.*, "The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[40] N. Pedano *et al.*, "The Cancer Genome Atlas Low Grade Glioma Collection (TCGA-LGG) (Version 3) [Data set]," The Cancer Imaging Archive, 2016, doi:10.7937/K9/TCIA.2016.L4LTD3TK.

[41] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[42] M. Buda, "brain-segmentation-pytorch," GitHub. https://github.com/mateuszbuda/brain-segmentation-pytorch

[43] P. Rao, N. Mathur, and A. Kalyanpur, "Utilization of teleradiology by intensive care units: A cohort study," *Indian J. Crit. Care Med.*, vol. 28, no. 1, pp. 20–25, 2024.

[44] K. Winstein, A. Sivaraman, and H. Balakrishnan, "Stochastic forecasts achieve high throughput and low delay over cellular networks," in *Proc. USENIX NSDI*, 2013, pp. 459–471.

[45] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Syst. Tech. J.*, vol. 39, no. 5, pp. 1253–1265, 1960.

[46] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell Syst. Tech. J.*, vol. 42, no. 5, pp. 1977–1997, 1963.

[47] A. Mehrtash, W. M. Wells III, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE Trans. Med. Imaging*, vol. 39, no. 12, pp. 3868–3878, 2020.

[48] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Process. Lett.*, vol. 4, no. 11, pp. 317–320, 1997.

[49] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.

[50] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.

[51] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979.

[52] V. E. Guarino *et al.*, "Better than average: Spatially-aware aggregation of segmentation uncertainty improves downstream performance," arXiv:2603.29941, 2026.

[53] A. Hiras *et al.*, "Task-specific quality gating for retinal optical coherence tomography B-scans: Learned representations over scalar metrics in choroid segmentation," medRxiv, 2026, doi:10.64898/2026.09.14.26363079.

[54] P. Steinmetz, F. Frouin, V. Morard, and I. Buvat, "Failure detection in biomedical image classification under realistic distribution shifts: Insights from a large-scale evaluation," medRxiv, v2, 2026, doi:10.64898/2026.05.04.26350496.

[55] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Hum. Factors*, vol. 39, no. 2, pp. 230–253, 1997.

[56] K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: A systematic review of frequency, effect mediators, and mitigators," *J. Am. Med. Inform. Assoc.*, vol. 19, no. 1, pp. 121–127, 2012.

Supplementary material: Does a Brain-Tumour Segmentation Model Notice When a Weak Telemedicine Link Has Degraded Its Input?

Adrija Das
Independent Researcher
formerly Kahlert School of Computing, University of Utah
Salt Lake City, UT, USA
adrija.das.cs@gmail.com (corresponding author)

Dr. Manabendra Maiti
Professor, Dept. of Electronics and Communication Engineering
Techno International New Town, Block DG 1/1, Action Area I,
New Town, Kolkata, West Bengal 700156, India
dr.manabendra.maiti@tict.edu.in

S1. Models and training

The split (70 training, 10 validation, 30 test patients) is stratified by acquisition site with a fixed seed. Because it is stratified by site, test patients share acquisition sites with training patients: the test is internal, not external. Each U-Net used the released recipe (Adam, learning rate 10^{-4}, batch 16, 100 epochs, Dice loss, the released augmentation), and, as in the released pipeline, the first and last slice of each study are dropped. The five U-Nets reached their best validation Dice at epochs 45, 20, 43, 20 and 44 (Dice 0.801, 0.798, 0.806, 0.814, 0.791); the MC-dropout U-Net at epoch 98 (0.828), so it may have benefited from longer training. The U-Nets were trained on a laptop GPU (Apple M5, PyTorch MPS) and nnU-Net v2 (2-D configuration, default plans) on one NVIDIA T4; all evaluation ran on two T4 GPUs. nnU-Net prediction used its own preprocessing, our 8 test-time transforms applied to its input, and no mirroring of its own. All models were trained without access to the test patients.

Which prediction is scored. Dice is always computed from each model's deterministic own-confidence map: one pass for U-Net, the member mean for the ensemble, dropout off for MC dropout, and the identity pass for nnU-Net. TTA and the other passes are used only for uncertainty. U-Net with TTA and U-Net with degradation-aware TTA therefore share their prediction, Dice and calibration, and differ only in their uncertainty; Tables S1 and S4 show them once. When a model predicts no tumour, its confidence in the predicted tumour is undefined; we set it to 0, the least confident case, which makes confidence look more responsive to failure, not less.

Hard patients. Two test patients score low before any transfer. TCGA_DU_7013 has clean Dice 0.064 (U-Net), 0.057 (ensemble), 0.000 (MC dropout) and 0.458 (nnU-Net); TCGA_HT_8105 has 0.462, 0.671, 0.228 and 0.810. On the slice of TCGA_DU_7013 with the largest labelled area, the label outlines a faint FLAIR hyperintensity and we saw no preprocessing artefact, so we kept the patient. The U-Net models almost entirely miss the lesion, so their Dice cannot fall further; nnU-Net finds part of it, and its 18 relative failures on this patient are losses of that partial detection. Section S7 repeats the failure-detection analysis without both patients.

S2. Calibration

Temperatures were fitted on the clean validation patients only, on logits: the pre-sigmoid output of each U-Net (read before the final sigmoid), the difference of nnU-Net's two class logits after its own resampling to the original grid, and the log-odds of the member mean for the ensemble. The negative log-likelihood was computed in the dilated tumour region used for ECE, and T was optimized continuously over $[0.05, 200]$. The temperatures were 6.34 (U-Net), 3.29 (ensemble mean), 9.12 (MC dropout) and 10.03 (nnU-Net); none reached a bound. Temperature scaling is monotone, so it changes calibration but no failure-detection result. For nnU-Net, the temperature that is optimal on the validation patients over-corrects clean test scans (confidence minus accuracy -0.138 near the tumour) yet leaves failures over-confident (0.099). Whole-image ECE is near zero for every model before scaling because background voxels dominate, which is why the tumour-region ECE is the main measure. After scaling it rises to 0.09–0.19 (Table S1): a temperature fitted in the tumour region also lowers the confidence of the many correctly classified background voxels, so one temperature does not serve both regions.

Table S1: Calibration near the tumour (ECE, ACE) before and after temperature scaling (T), and over the whole image, for clean studies, all degraded studies delivered in time and failures (Dice lost ≥ 0.10). U-Net with degradation-aware TTA shares U-Net's prediction and is not repeated.

Model	Studies	ECE	ECE, T	ACE	ACE, T	ECE, image	ECE, image, T	ACE, image
U-Net	clean (30)	0.168	0.047	0.163	0.085	0.0033	0.1780	0.151
	degraded (1314)	0.189	0.062	0.172	0.096	0.0037	0.1763	0.159
	failures (111)	0.334	0.176	0.195	0.141	0.0057	0.1671	0.172
Ensemble	clean (30)	0.104	0.030	0.129	0.043	0.0006	0.1002	0.081
	degraded (1314)	0.118	0.034	0.132	0.045	0.0008	0.0993	0.072
	failures (98)	0.245	0.135	0.195	0.115	0.0027	0.0949	0.079
MC dropout	clean (30)	0.154	0.036	0.158	0.049	0.0031	0.1730	0.147
	degraded (1314)	0.182	0.056	0.167	0.065	0.0037	0.1714	0.155
	failures (148)	0.312	0.162	0.195	0.127	0.0059	0.1670	0.174
nnU-Net	clean (30)	0.131	0.149	0.205	0.211	0.0026	0.1885	0.196
	degraded (1314)	0.158	0.127	0.207	0.175	0.0032	0.1882	0.196
	failures (137)	0.359	0.099	0.233	0.267	0.0063	0.1860	0.208

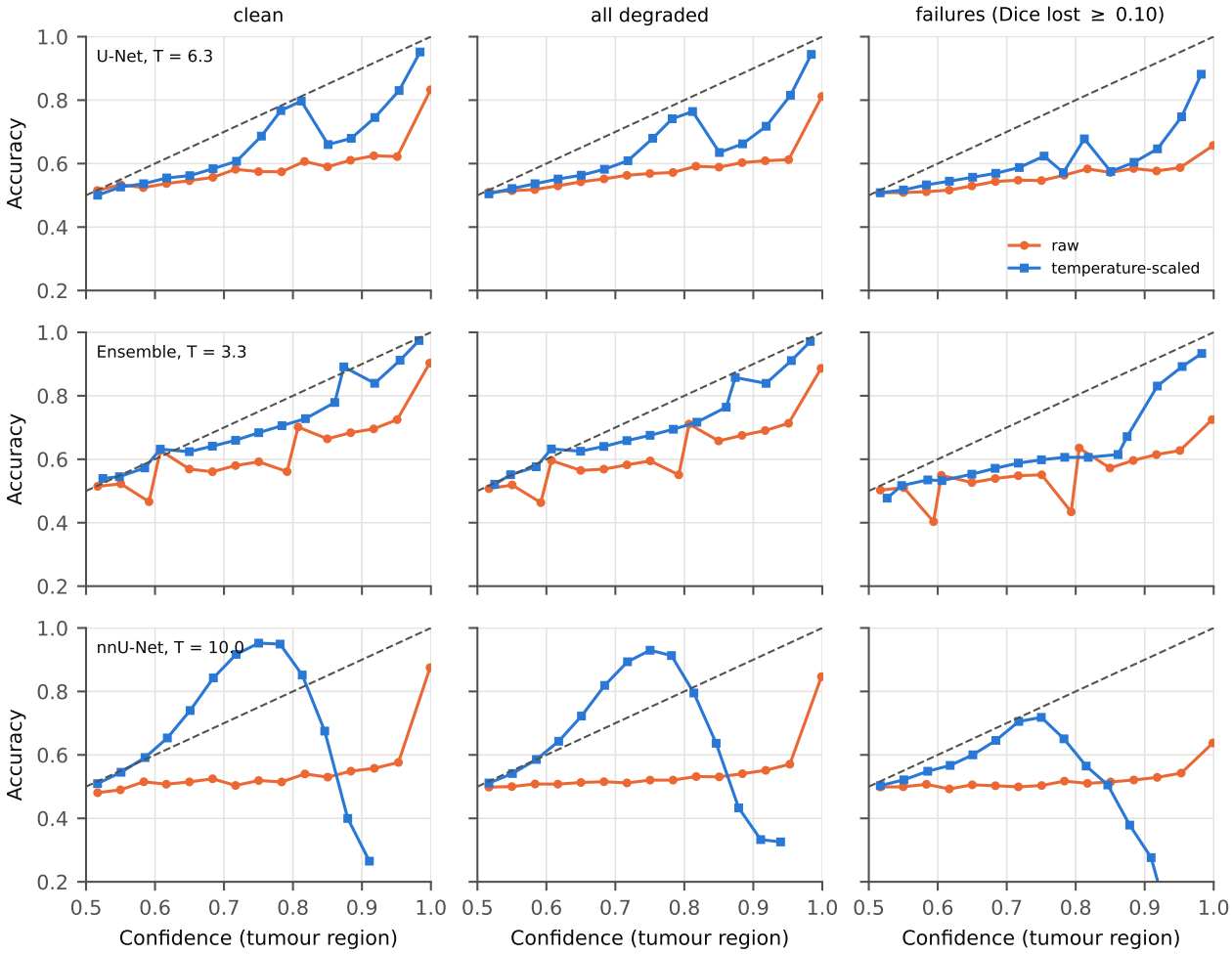


Figure S1: Reliability near the tumour before (orange) and after (blue) temperature scaling for U-Net, the ensemble and nnU-Net, on clean studies, all degraded studies and failures. Bins with fewer than 50 voxels are omitted.

S3. Failure detection: all scores, both risk views, and thresholds

Table S2 lists every warning for every model, including U-Net with degradation-aware test-time augmentation, with AURC for transmission risk and AUROC for an absolute failure (Dice < 0.7). Absolute failures were 251 studies from 19 patients for U-Net and 156 from 20 for nnU-Net. Fig. S2 shows AUROC for relative failures defined by Dice losses of 0.05–0.20.

The uncertainty ratio gave AURC 0.110, 0.109 and 0.109 for U-Net at $\tau = 0.3$, 0.5 and 0.7 bit, and 0.095 to 0.160 for MC dropout. QU-BraTS scores on degraded studies were 0.832 (U-Net) and 0.848 (nnU-Net).

Table S2: Failure detection, transmission view: AURC with risk = Dice lost (lower is better; best per model in bold), with AUROC for an absolute failure (Dice < 0.7) in brackets.

Warning score	U-Net	U-Net, deg. TTA	Ensemble	MC dropout	nnU-Net
Mean confidence in predicted tumour	0.017 (0.76)	0.017 (0.76)	0.010 (0.98)	0.023 (0.83)	0.013 (0.84)
Pairwise Dice disagreement	0.012 (0.97)	0.015 (0.95)	0.009 (0.98)	0.013 (0.98)	0.012 (0.92)
Uncertainty ratio, $H > 0.5$ bit	0.012 (0.96)	0.015 (0.94)	0.009 (0.97)	0.014 (0.95)	0.012 (0.93)
Uncertainty ratio, $H > 0.3$ bit	0.012 (0.96)	0.015 (0.94)	0.010 (0.97)	0.014 (0.99)	0.012 (0.92)
Uncertainty ratio, $H > 0.7$ bit	0.013 (0.95)	0.014 (0.96)	0.009 (0.97)	0.013 (0.90)	0.012 (0.91)
Mean entropy in predicted tumour	0.014 (0.97)	0.016 (0.95)	0.010 (0.99)	0.016 (0.85)	0.018 (0.91)
Boundary-band entropy	0.017 (0.97)	0.019 (0.95)	0.012 (0.98)	0.018 (0.84)	0.021 (0.90)
Volume change vs. mean prediction	0.020 (0.83)	0.017 (0.85)	n/a	0.019 (0.78)	0.023 (0.82)
Volume CV across passes	0.017 (0.89)	0.019 (0.87)	0.011 (0.95)	0.014 (0.94)	0.014 (0.92)
Lost-strip fraction	0.014 (0.59)	0.014 (0.59)	0.012 (0.57)	0.016 (0.62)	0.017 (0.63)
Metadata regression	0.010 (0.46)	0.010 (0.46)	0.008 (0.46)	0.011 (0.50)	0.014 (0.51)
Receiver image-quality regression	0.010 (0.44)	0.010 (0.44)	0.009 (0.38)	0.011 (0.53)	0.013 (0.55)
Combined regression	**0.008** (0.96)	**0.008** (0.95)	**0.007** (0.98)	**0.010** (0.97)	**0.010** (0.93)
Oracle / random	0.004 / 0.031	0.004 / 0.031	0.003 / 0.027	0.004 / 0.037	0.004 / 0.036
Failures: studies / patients	111 / 22	111 / 22	98 / 22	148 / 24	137 / 25

Table S3: Failure detection over all degraded studies with 95% patient-cluster bootstrap intervals.

Model	Warning	AURC, 1−Dice	AURC, Dice lost	AUROC, lost ≥0.10	AUROC, Dice <0.7
U-Net	Mean confidence in predicted tumour	0.141 [0.108, 0.191]	0.017 [0.014, 0.020]	0.77 [0.69, 0.84]	0.76 [0.66, 0.86]
	Pairwise Dice disagreement	0.109 [0.091, 0.142]	0.012 [0.010, 0.015]	0.86 [0.78, 0.93]	0.97 [0.94, 0.99]
	Uncertainty ratio, $H > 0.5$ bit	0.109 [0.090, 0.143]	0.012 [0.010, 0.015]	0.87 [0.79, 0.93]	0.96 [0.92, 0.99]
	Uncertainty ratio, $H > 0.3$ bit	0.110 [0.088, 0.143]	0.012 [0.010, 0.015]	0.87 [0.79, 0.93]	0.96 [0.92, 0.99]
	Uncertainty ratio, $H > 0.7$ bit	0.109 [0.089, 0.142]	0.013 [0.010, 0.015]	0.87 [0.79, 0.93]	0.95 [0.92, 0.99]
	Mean entropy in predicted tumour	0.111 [0.090, 0.139]	0.014 [0.011, 0.017]	0.82 [0.74, 0.90]	0.97 [0.93, 0.99]
	Boundary-band entropy	0.124 [0.099, 0.162]	0.017 [0.013, 0.021]	0.76 [0.66, 0.84]	0.97 [0.93, 0.99]
	Volume change vs. mean prediction	0.132 [0.104, 0.177]	0.020 [0.015, 0.027]	0.72 [0.63, 0.80]	0.83 [0.70, 0.93]
	Volume CV across passes	0.122 [0.097, 0.156]	0.017 [0.013, 0.022]	0.74 [0.66, 0.82]	0.89 [0.79, 0.95]
	Lost-strip fraction	0.189 [0.133, 0.276]	0.014 [0.011, 0.020]	0.80 [0.76, 0.84]	0.59 [0.54, 0.69]
	Metadata regression	0.214 [0.150, 0.294]	0.010 [0.007, 0.013]	0.85 [0.82, 0.88]	0.46 [0.36, 0.60]
	Receiver image-quality regression	0.250 [0.155, 0.352]	0.010 [0.008, 0.012]	0.91 [0.89, 0.93]	0.44 [0.32, 0.71]
	Combined regression	0.111 [0.090, 0.141]	0.008 [0.006, 0.010]	0.92 [0.85, 0.96]	0.96 [0.93, 0.99]
U-Net, deg. TTA	Mean confidence in predicted tumour	0.141 [0.106, 0.191]	0.017 [0.014, 0.020]	0.77 [0.69, 0.84]	0.76 [0.66, 0.86]
	Pairwise Dice disagreement	0.112 [0.091, 0.142]	0.015 [0.012, 0.019]	0.82 [0.74, 0.89]	0.95 [0.89, 0.99]
	Uncertainty ratio, $H > 0.5$ bit	0.113 [0.090, 0.147]	0.015 [0.012, 0.020]	0.82 [0.74, 0.89]	0.94 [0.89, 0.98]
	Uncertainty ratio, $H > 0.3$ bit	0.112 [0.089, 0.141]	0.015 [0.012, 0.019]	0.82 [0.75, 0.89]	0.94 [0.88, 0.98]
	Uncertainty ratio, $H > 0.7$ bit	0.110 [0.092, 0.142]	0.014 [0.011, 0.018]	0.84 [0.76, 0.90]	0.96 [0.92, 0.99]
	Mean entropy in predicted tumour	0.115 [0.095, 0.150]	0.016 [0.013, 0.022]	0.76 [0.67, 0.84]	0.95 [0.89, 0.99]
	Boundary-band entropy	0.125 [0.103, 0.156]	0.019 [0.016, 0.023]	0.73 [0.65, 0.81]	0.95 [0.90, 0.98]
	Volume change vs. mean prediction	0.139 [0.108, 0.184]	0.017 [0.014, 0.022]	0.72 [0.64, 0.79]	0.85 [0.77, 0.91]
	Volume CV across passes	0.136 [0.108, 0.176]	0.019 [0.014, 0.025]	0.74 [0.64, 0.81]	0.87 [0.72, 0.95]
	Lost-strip fraction	0.189 [0.139, 0.273]	0.014 [0.011, 0.021]	0.80 [0.76, 0.84]	0.59 [0.54, 0.68]
	Metadata regression	0.214 [0.147, 0.302]	0.010 [0.007, 0.012]	0.85 [0.82, 0.88]	0.46 [0.34, 0.60]
	Receiver image-quality regression	0.250 [0.156, 0.344]	0.010 [0.008, 0.013]	0.91 [0.89, 0.93]	0.44 [0.32, 0.69]
	Combined regression	0.111 [0.089, 0.145]	0.008 [0.006, 0.011]	0.91 [0.86, 0.95]	0.95 [0.90, 0.99]
Ensemble	Mean confidence in predicted tumour	0.091 [0.074, 0.115]	0.010 [0.008, 0.013]	0.87 [0.79, 0.94]	0.98 [0.96, 0.99]
	Pairwise Dice disagreement	0.095 [0.078, 0.122]	0.009 [0.007, 0.012]	0.89 [0.81, 0.96]	0.98 [0.96, 0.99]
	Uncertainty ratio, $H > 0.5$ bit	0.096 [0.080, 0.123]	0.009 [0.008, 0.012]	0.89 [0.81, 0.96]	0.97 [0.94, 0.99]
	Uncertainty ratio, $H > 0.3$ bit	0.096 [0.080, 0.120]	0.010 [0.008, 0.012]	0.89 [0.79, 0.96]	0.97 [0.94, 0.99]
	Uncertainty ratio, $H > 0.7$ bit	0.097 [0.078, 0.125]	0.009 [0.007, 0.012]	0.89 [0.81, 0.96]	0.97 [0.95, 0.99]
	Mean entropy in predicted tumour	0.092 [0.076, 0.118]	0.010 [0.008, 0.013]	0.85 [0.76, 0.93]	0.99 [0.97, 1.00]
	Boundary-band entropy	0.098 [0.079, 0.127]	0.012 [0.010, 0.016]	0.81 [0.71, 0.90]	0.98 [0.96, 0.99]
	Volume change vs. mean prediction	n/a	n/a	n/a	n/a
	Volume CV across passes	0.101 [0.084, 0.127]	0.011 [0.009, 0.014]	0.83 [0.73, 0.91]	0.95 [0.90, 0.98]
	Lost-strip fraction	0.169 [0.123, 0.254]	0.012 [0.009, 0.017]	0.82 [0.78, 0.85]	0.57 [0.52, 0.66]
	Metadata regression	0.188 [0.128, 0.262]	0.008 [0.006, 0.010]	0.87 [0.85, 0.89]	0.46 [0.34, 0.63]
	Receiver image-quality regression	0.234 [0.134, 0.350]	0.009 [0.007, 0.011]	0.91 [0.89, 0.94]	0.38 [0.27, 0.66]

Table S3: Failure detection over all degraded studies with 95% patient-cluster bootstrap intervals.

Model	Warning	AURC, 1−Dice	AURC, Dice lost	AUROC, lost ≥0.10	AUROC, Dice <0.7
	Combined regression	0.094 [0.077, 0.122]	0.007 [0.005, 0.009]	0.91 [0.84, 0.95]	0.98 [0.96, 0.99]
MC dropout	Mean confidence in predicted tumour	0.120 [0.094, 0.160]	0.023 [0.018, 0.027]	0.71 [0.62, 0.78]	0.83 [0.74, 0.90]
	Pairwise Dice disagreement	0.094 [0.079, 0.119]	0.013 [0.010, 0.017]	0.85 [0.76, 0.94]	0.98 [0.96, 0.99]
	Uncertainty ratio, $H > 0.5$ bit	0.125 [0.079, 0.192]	0.014 [0.011, 0.018]	0.85 [0.77, 0.93]	0.95 [0.89, 0.99]
	Uncertainty ratio, $H > 0.3$ bit	0.095 [0.077, 0.123]	0.014 [0.011, 0.018]	0.85 [0.76, 0.94]	0.99 [0.97, 0.99]
	Uncertainty ratio, $H > 0.7$ bit	0.160 [0.080, 0.262]	0.013 [0.011, 0.017]	0.86 [0.78, 0.93]	0.90 [0.77, 0.99]
	Mean entropy in predicted tumour	0.188 [0.084, 0.329]	0.016 [0.010, 0.030]	0.85 [0.78, 0.92]	0.85 [0.64, 0.99]
	Boundary-band entropy	0.194 [0.090, 0.321]	0.018 [0.012, 0.033]	0.82 [0.75, 0.89]	0.84 [0.65, 0.97]
	Volume change vs. mean prediction	0.194 [0.095, 0.329]	0.019 [0.013, 0.031]	0.79 [0.73, 0.85]	0.78 [0.60, 0.91]
	Volume CV across passes	0.101 [0.081, 0.129]	0.014 [0.011, 0.019]	0.82 [0.73, 0.90]	0.94 [0.90, 0.98]
	Lost-strip fraction	0.171 [0.114, 0.277]	0.016 [0.012, 0.025]	0.80 [0.76, 0.84]	0.62 [0.56, 0.78]
	Metadata regression	0.198 [0.133, 0.288]	0.011 [0.008, 0.015]	0.86 [0.82, 0.89]	0.50 [0.37, 0.64]
	Receiver image-quality regression	0.213 [0.127, 0.307]	0.011 [0.008, 0.015]	0.92 [0.89, 0.94]	0.53 [0.40, 0.79]
	Combined regression	0.116 [0.084, 0.165]	0.010 [0.007, 0.014]	0.91 [0.83, 0.96]	0.97 [0.94, 0.99]
nnU-Net	Mean confidence in predicted tumour	0.097 [0.079, 0.121]	0.013 [0.010, 0.016]	0.87 [0.81, 0.91]	0.84 [0.72, 0.97]
	Pairwise Dice disagreement	0.093 [0.077, 0.116]	0.012 [0.010, 0.015]	0.90 [0.86, 0.93]	0.92 [0.84, 0.98]
	Uncertainty ratio, $H > 0.5$ bit	0.093 [0.076, 0.116]	0.012 [0.010, 0.015]	0.90 [0.86, 0.93]	0.93 [0.86, 0.98]
	Uncertainty ratio, $H > 0.3$ bit	0.093 [0.078, 0.119]	0.012 [0.010, 0.015]	0.90 [0.86, 0.93]	0.92 [0.84, 0.98]
	Uncertainty ratio, $H > 0.7$ bit	0.093 [0.077, 0.114]	0.012 [0.010, 0.015]	0.90 [0.86, 0.93]	0.91 [0.82, 0.98]
	Mean entropy in predicted tumour	0.102 [0.078, 0.134]	0.018 [0.010, 0.029]	0.87 [0.82, 0.92]	0.91 [0.86, 0.95]
	Boundary-band entropy	0.109 [0.083, 0.140]	0.021 [0.014, 0.032]	0.82 [0.77, 0.87]	0.90 [0.84, 0.94]
	Volume change vs. mean prediction	0.130 [0.106, 0.162]	0.023 [0.016, 0.033]	0.78 [0.71, 0.83]	0.82 [0.70, 0.94]
	Volume CV across passes	0.100 [0.083, 0.125]	0.014 [0.011, 0.017]	0.88 [0.83, 0.92]	0.92 [0.85, 0.97]
	Lost-strip fraction	0.150 [0.113, 0.198]	0.017 [0.013, 0.023]	0.76 [0.73, 0.79]	0.63 [0.55, 0.79]
	Metadata regression	0.170 [0.118, 0.229]	0.014 [0.008, 0.020]	0.79 [0.73, 0.87]	0.51 [0.41, 0.72]
	Receiver image-quality regression	0.177 [0.122, 0.233]	0.013 [0.008, 0.018]	0.88 [0.84, 0.93]	0.55 [0.43, 0.89]
	Combined regression	0.093 [0.077, 0.114]	0.010 [0.005, 0.017]	0.94 [0.89, 0.97]	0.93 [0.86, 0.99]

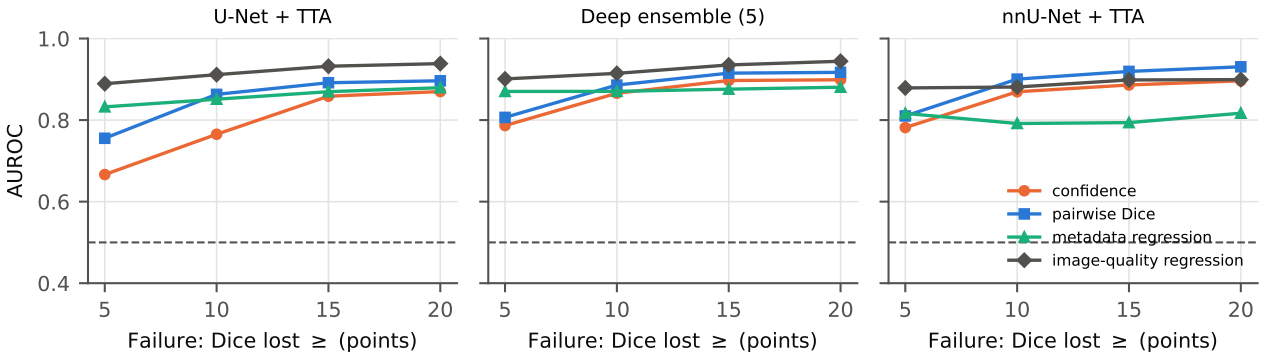


Figure S2: AUROC of four warnings against the threshold that defines a relative failure.

S4. Examples

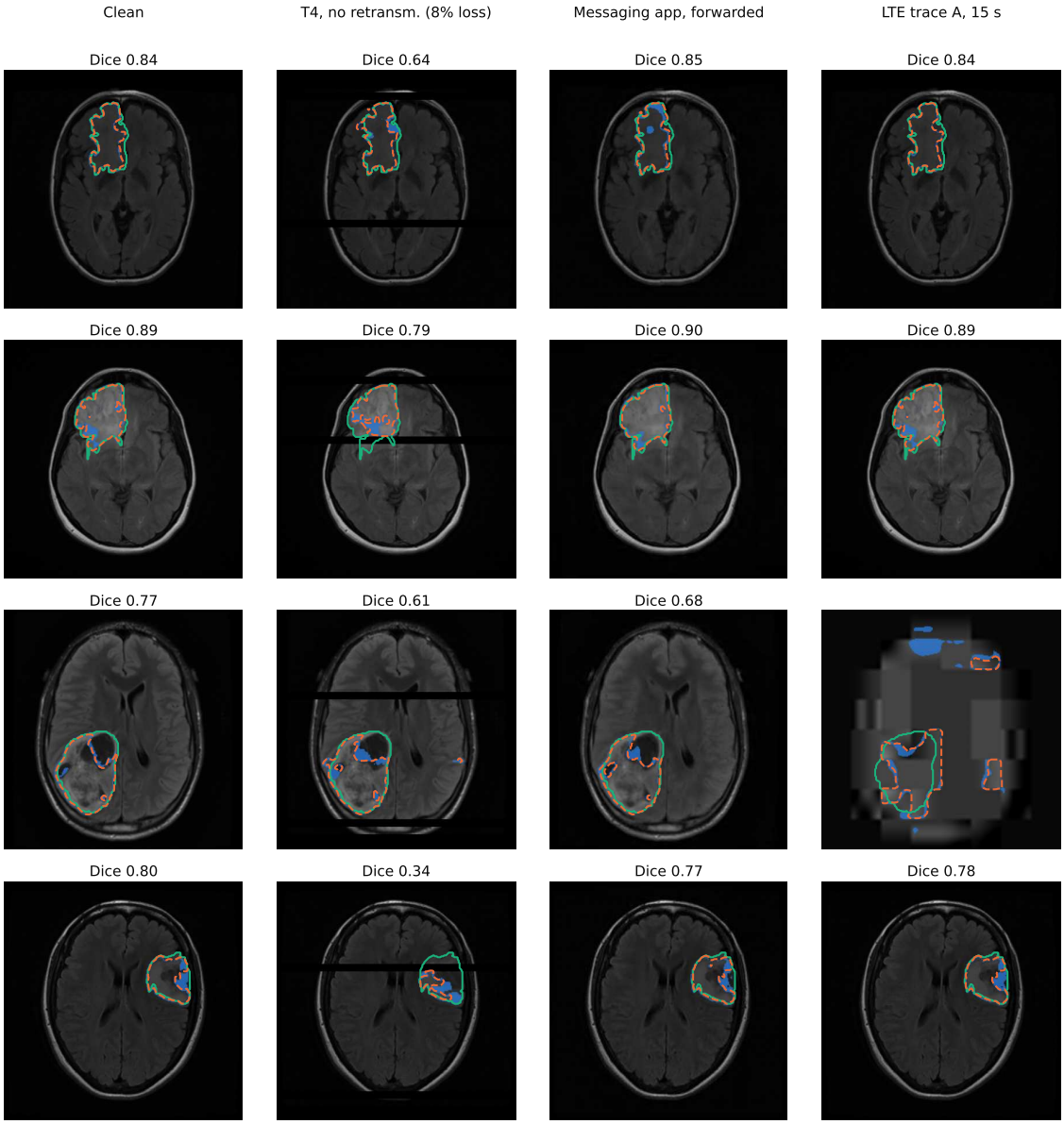


Figure S3: Four test patients selected by a fixed rule (the two largest drops under any condition and two at the median and lower quartile of clean Dice: TCGA_DU_7300_19910814, TCGA_HT_7874_19950902, TCGA_CS_6666_20011109, TCGA_HT_A616_19991226), slice with the largest tumour area. Received FLAIR with ground truth (green), U-Net prediction (orange, dashed) and voxels with TTA entropy above 0.5 bit (blue). The fourth column shows a trace window that delivered nothing before the deadline (third row), after which the sender's floor setting ($\times 0.25$, JPEG quality 5) arrived late.

S5. Accuracy by condition

Table S4: Mean Dice per condition and model, with failures (Dice lost ≥ 0.10) in brackets (* decrease vs. clean, Wilcoxon, Holm $p < 0.05$; † equivalent within ± 0.01, TOST, Holm $p < 0.05$; both marks: a consistent decrease smaller than the margin). Conditions with random draws (strip loss, real-time tiers, traces) were run twice; studies that missed the deadline are excluded (Table S9). T2 without retransmission is the 1% strip-loss row. U-Net with degradation-aware TTA has the same Dice as U-Net and is not repeated.

Condition	U-Net	Ensemble	MC dropout	nnU-Net
Clean	0.817	0.837	0.830	0.865
JPEG q 75	0.816	0.835	0.826	0.862*
JPEG q 50	0.817	0.837	0.828	0.856* (1)
JPEG q 30	0.813 (1)	0.834	0.816* (1)	0.856* (1)
JPEG q 20	0.805*	0.827*	0.815* (1)	0.848*
JPEG q 10	0.799* (1)	0.821 (1)	0.798* (2)	0.827* (3)
JPEG q 5	0.664* (12)	0.713* (10)	0.671* (15)	0.693* (18)
1% strips, zero-filled (= T2, no retransmission)	0.802* (1)	0.827*	0.813* (2)	0.850* (2)
2% strips, zero-filled	0.798* (3)	0.823*	0.802* (3)	0.842* (4)
5% strips, zero-filled	0.746* (14)	0.775* (12)	0.736* (20)	0.789* (12)
10% strips, zero-filled	0.662* (34)	0.696* (28)	0.651* (39)	0.694* (34)
5% strips, concealed	0.815	0.835	0.827*	0.859*
10% strips, concealed	0.815	0.833	0.825*	0.851* (1)
5% strips, bursty	0.735* (14)	0.751* (19)	0.712* (28)	0.754* (24)
5% strips, bursty, concealed	0.793* (3)	0.810* (3)	0.810* (2)	0.819* (6)
Messaging app, sent once	0.812	0.833	0.826	0.861
Messaging app, forwarded	0.812	0.832	0.827	0.861
T1, reliable, 60 s	0.817†	0.837†	0.830†	0.865†
T1, no retransmission (0.1% loss)	0.817	0.836	0.829	0.863*
T2, reliable, 60 s	0.817†	0.837†	0.830†	0.865†
T3, reliable, 60 s	0.816†	0.837†	0.831†	0.864†
T3, no retransmission (3% loss)	0.786* (6)	0.813* (2)	0.785* (6)	0.828* (6)
T4, reliable, 60 s	0.819	0.838†	0.830†	0.861 (1)
T4, no retransmission (8% loss)	0.709* (22)	0.739* (23)	0.694* (29)	0.730* (23)
T3, reliable, 15 s	0.815†	0.836*†	0.827†	0.860*
T3, reliable, 300 s	0.817†	0.837†	0.830†	0.865†
T4, reliable, 15 s	0.807	0.830	0.812*	0.838* (1)
T4, reliable, 300 s	0.815†	0.835†	0.830†	0.864†
LTE trace A, 60 s	0.817†	0.837†	0.830†	0.865†
LTE trace A, 15 s	0.815†	0.836†	0.831†	0.864†
LTE trace B, 15 s	0.817†	0.836†	0.830†	0.865†

Table S5: Codec setting chosen per test patient for rate-controlled conditions (U-Net run; seed 0 for traces). "late": no setting fitted the budget, so the floor setting was sent and missed the deadline.

Patient	T3, 60 s	T4, 60 s	T3, 15 s	T4, 15 s	LTE trace A, 15 s
TCGA_CS_4942	lossless	q 90	q 90	q 40	lossless
TCGA_CS_4943	lossless	q 90	q 85	q 30	lossless
TCGA_CS_5396	lossless	q 90	q 85	q 50, $\times$0.75	lossless
TCGA_CS_6665	lossless	q 90	q 85	q 55, $\times$0.75	lossless
TCGA_CS_6666	lossless	q 85	q 80	q 40, $\times$0.75	–
TCGA_DU_5853	q 95	q 80	q 70	q 25, $\times$0.75	q 90
TCGA_DU_5855	lossless	q 90	q 90	q 40	q 90
TCGA_DU_5874	q 95	q 80	q 70	q 65, $\times$0.5	q 95
TCGA_DU_6400	q 90	q 45	q 30	q 50, $\times$0.375	q 85
TCGA_DU_6404	q 90	q 60	q 45	q 35, $\times$0.5	q 85
TCGA_DU_6407	q 90	q 50	q 35	q 20, $\times$0.5	q 85
TCGA_DU_6408	q 90	q 65	q 50	q 35, $\times$0.5	q 80
TCGA_DU_7010	q 90	q 45	q 30	q 45, $\times$0.375	q 90
TCGA_DU_7013	q 85	q 30	q 25	q 20, $\times$0.5	q 85
TCGA_DU_7300	lossless	q 85	q 85	q 40, $\times$0.75	q 95
TCGA_DU_8163	q 95	q 80	q 70	q 65, $\times$0.5	q 95
TCGA_DU_8168	q 95	q 80	q 75	q 25, $\times$0.75	q 95
TCGA_FG_5962	q 95	q 65	q 50	q 35, $\times$0.5	q 50
TCGA_FG_5964	q 95	q 80	q 70	q 65, $\times$0.5	q 95
TCGA_FG_6691	q 95	q 75	q 65	q 55, $\times$0.5	q 90
TCGA_FG_A4MU	q 95	q 80	q 70	q 50, $\times$0.5	q 95
TCGA_HT_7475	q 95	q 75	q 65	q 55, $\times$0.5	q 95
TCGA_HT_7616	lossless	q 80	q 75	q 60, $\times$0.5	lossless
TCGA_HT_7684	lossless	q 90	q 85	q 35, $\times$0.75	q 95
TCGA_HT_7874	lossless	q 95	q 90	q 35	lossless
TCGA_HT_7877	q 95	q 85	q 80	q 30, $\times$0.75	q 95
TCGA_HT_8105	q 95	q 75	q 70	q 25, $\times$0.75	q 95
TCGA_HT_8107	lossless	q 90	q 90	q 35	lossless
TCGA_HT_8111	lossless	q 90	q 85	q 55, $\times$0.75	lossless
TCGA_HT_A616	lossless	q 85	q 85	q 45, $\times$0.75	q 95

Table S6: Failures (Dice lost ≥ 0.10) per test patient and model, with the patient's clean Dice for U-Net.

Patient	U-Net	U-Net, deg. TTA	Ensemble	MC dropout	nnU-Net	Clean Dice (U-Net)
TCGA_CS_4942	2	2	1	3	2	0.92
TCGA_CS_4943	0	0	0	0	2	0.91
TCGA_CS_5396	4	4	7	8	2	0.65
TCGA_CS_6665	0	0	1	3	2	0.54
TCGA_CS_6666	6	6	7	8	10	0.77
TCGA_DU_5853	11	11	9	7	6	0.93
TCGA_DU_5855	3	3	3	5	6	0.94
TCGA_DU_5874	9	9	8	8	4	0.86
TCGA_DU_6400	0	0	0	0	0	0.95
TCGA_DU_6404	9	9	6	9	7	0.95
TCGA_DU_6407	2	2	2	3	2	0.92
TCGA_DU_6408	1	1	0	1	0	0.94
TCGA_DU_7010	0	0	0	3	4	0.76
TCGA_DU_7013	0	0	0	0	18	0.06
TCGA_DU_7300	8	8	5	8	12	0.84
TCGA_DU_8163	5	5	4	9	8	0.90
TCGA_DU_8168	8	8	3	9	3	0.87
TCGA_FG_5962	2	2	2	7	0	0.85
TCGA_FG_5964	4	4	3	6	3	0.83
TCGA_FG_6691	0	0	1	0	2	0.94
TCGA_FG_A4MU	2	2	3	4	2	0.91
TCGA_HT_7475	0	0	0	0	0	0.90
TCGA_HT_7616	1	1	1	2	2	0.75
TCGA_HT_7684	0	0	0	0	0	0.96
TCGA_HT_7874	1	1	0	3	2	0.89
TCGA_HT_7877	10	10	8	13	7	0.73
TCGA_HT_8105	7	7	7	3	14	0.46
TCGA_HT_8107	1	1	3	4	5	0.85
TCGA_HT_8111	2	2	4	5	4	0.94
TCGA_HT_A616	13	13	10	17	8	0.80

S6. Simulator validation

Table S7: Checks run on synthetic phantoms before any patient data (35 of 35 passed). Checks 8 and 22 are negative controls: deliberately broken operations that the checks must detect.

#	Check	Detail	Passed
1	JPEG bytes decrease monotonically as quality drops	11,741 B at q 95; 662 B at q 5	yes
2	JPEG PSNR decreases monotonically as quality drops	46.5 dB at q 95; 28.5 dB at q 5	yes
3	Lossless path is exact		yes
4	Down- and up-sampling by 0.75: centroid shift $<$ 0.1 px and area error $<$ 3%	shift 0.024 px, area error 0.18%	yes
5	Down- and up-sampling by 0.5: centroid shift $<$ 0.1 px and area error $<$ 3%	shift 0.005 px, area error 0.07%	yes
6	Down- and up-sampling by 0.375: centroid shift $<$ 0.1 px and area error $<$ 3%	shift 0.096 px, area error 0.00%	yes
7	Down- and up-sampling by 0.25: centroid shift $<$ 0.1 px and area error $<$ 3%	shift 0.000 px, area error 0.21%	yes
8	Negative control: naive 2:1 decimation is detected as shifted ($>$0.25 px)	shift 0.707 px	yes
9	Strip loss rate matches p within 3 standard errors	0.0511 at $p = 0.05$	yes
10	Lost strips are zero-filled, kept strips unchanged		yes
11	Loss probability of a multi-packet strip is $1 - (1 - p)^k$		yes
12	Mathis throughput matches a hand calculation	950 kb/s	yes
13	TCP throughput decreases from T1 to T4	9012, 950, 274, 84 kb/s	yes
14	Rate control, budget 2.00$\times$ raw size: the chosen setting fits, and its size matches the transmitted study	full resolution, lossless, 274,926 B	yes
15	Rate control, budget 0.30$\times$ raw size: the chosen setting fits, and its size matches the transmitted study	full resolution, lossless, 274,926 B	yes
16	Rate control, budget 0.10$\times$ raw size: the chosen setting fits, and its size matches the transmitted study	full resolution, JPEG q 85, 105,912 B	yes
17	Rate control, budget 0.03$\times$ raw size: the chosen setting fits, and its size matches the transmitted study	full resolution, JPEG q 30, 33,735 B	yes
18	Rate control, budget 0.01$\times$ raw size: the chosen setting fits, and its size matches the transmitted study	scale 0.375, JPEG q 30, 11,241 B	yes
19	Generous budget gives lossless delivery		yes
20	Mask survives preprocessing round trip (Dice > 0.97)	Dice 1.0000	yes
21	Round-trip centroid shift $<$ 0.5 px	shift 0.000 px	yes
22	Negative control: skipping the un-padding step is detected		yes
23	TTA mean equals the identity pass for a pixel-wise model	max. difference 1.2×10^{-7}	yes
24	ECE of a perfectly calibrated predictor is $\approx$0	ECE 0.0020	yes
25	ECE of a 95%-confident, 75%-accurate predictor is $\approx$0.20	ECE 0.2002	yes
26	Dice of two half-overlapping bands is 0.5		yes
27	Gilbert–Elliott loss: loss rate $\approx$5% and mean burst $\approx$4 strips	loss rate 0.0492, mean burst 3.99 strips	yes
28	Concealment restores a smooth image (error $\leq$ 2) where zero-fill does not	max. error 1 (concealed), 150 (zero-filled)	yes
29	Concealment leaves received strips unchanged		yes
30	Messaging app: forwarding again lowers fidelity	PSNR 37.3 dB once, 37.0 dB twice	yes
31	Adding concealment and bursty loss leaves independent zero-filled loss unchanged (identical image and size)		yes
32	Trace budget = packets in the window $\times$ MSS		yes
33	Trace budget grows with the deadline		yes
34	Trace mean rate matches its stated mean		yes
35	Fast slice-wise resize equals the released 4-D resize	max. difference 8.9×10^{-16}	yes

S7. Further analyses

Ensemble members. The ensemble's confidence map is the mean of five members, so it contains their disagreement. Table S8 scores each member's own confidence as a warning, which is the like-for-like comparison with the single U-Net (member 1 is that U-Net, so its row matches the U-Net confidence entry of Table II).

Table S8: Each ensemble member's own confidence as a warning: AURC for absolute risk and AUROC for Dice lost ≥ 0.10, over its own degraded studies, against the ensemble mean's confidence and pairwise Dice.

Model	Clean Dice	Failures	AURC (conf.)	AUROC (conf.)	Oracle / random
Member 1 (= U-Net)	0.817	111	0.141	0.77	0.098 / 0.211
Member 2	0.834	130	0.128	0.78	0.103 / 0.197
Member 3	0.820	127	0.127	0.73	0.094 / 0.207
Member 4	0.821	131	0.138	0.74	0.104 / 0.207
Member 5	0.828	140	0.154	0.74	0.098 / 0.206
Members, mean	0.824	128	0.137	0.75	0.099 / 0.206
Ensemble, confidence of mean	0.837	98	0.091	0.87	0.087 / 0.188
Ensemble, pairwise Dice	0.837	98	0.095	0.89	0.087 / 0.188

Late studies. Table S9 shows the trace conditions with and without the studies that missed the deadline.

Table S9: LTE-trace conditions: change in mean Dice from clean ($\times 100$) with failures in brackets, including / excluding the studies that missed the deadline. Trace A: AT&T LTE uplink; trace B: Verizon LTE uplink.

Condition	Late	U-Net	Ensemble	MC dropout	nnU-Net
LTE trace A, 60 s	1 of 60	−0.7 (1) / 0.0 (0)	−0.6 (1) / 0.0 (0)	−0.8 (1) / 0.0 (0)	−1.0 (1) / 0.0 (0)
LTE trace A, 15 s	5 of 60	−3.7 (5) / −0.2 (0)	−3.4 (5) / −0.1 (0)	−3.4 (5) / +0.1 (0)	−4.1 (5) / −0.1 (0)
LTE trace B, 15 s	0 of 60	0.0 (0) / 0.0 (0)	0.0 (0) / 0.0 (0)	0.0 (0) / 0.0 (0)	0.0 (0) / 0.0 (0)

Without the two hard patients. Table S10 repeats the failure-detection analysis without TCGA_DU_7013 and TCGA_HT_8105, refitting the regressions leave-one-patient-out on the remaining 28 patients.

Table S10: Failure detection without the two patients who are hard before transfer: AURC for absolute risk with AUROC for Dice lost ≥ 0.10 in brackets.

Warning score	U-Net	Ensemble	MC dropout	nnU-Net
Mean confidence in predicted tumour	0.123 (0.76)	0.086 (0.92)	0.107 (0.74)	0.093 (0.87)
Pairwise Dice disagreement	0.103 (0.91)	0.090 (0.93)	0.087 (0.91)	0.089 (0.90)
Metadata regression	0.165 (0.85)	0.154 (0.87)	0.137 (0.85)	0.135 (0.88)
Receiver image-quality regression	0.151 (0.91)	0.137 (0.91)	0.149 (0.92)	0.146 (0.91)
Combined regression	0.107 (0.95)	0.088 (0.94)	0.087 (0.95)	0.089 (0.96)
Oracle / random	0.092 / 0.173	0.082 / 0.156	0.081 / 0.157	0.081 / 0.148

Held-out degradation families. The metadata, image-quality and combined regressions were fitted (leave-one-patient-out) only on the JPEG sweep, independent zero-filled strip loss, the assumed tiers and the deadline variants, and tested on the families they never saw: the messaging app, concealment, bursty loss (zero-filled and concealed) and the LTE traces.

Table S11: AUROC for Dice lost ≥ 0.10 on the held-out families, with the AUROC of the same regression fitted on all families in brackets. Confidence and pairwise Dice need no fitting.

Warning (AUROC, Dice lost ≥0.10)	U-Net	Ensemble	MC dropout	nnU-Net
Metadata regression	0.60 (0.69)	0.64 (0.69)	0.69 (0.73)	0.67 (0.67)
Image-quality regression	0.89 (0.88)	0.90 (0.90)	0.92 (0.92)	0.84 (0.84)
Combined regression	0.90 (0.93)	0.89 (0.91)	0.92 (0.93)	0.87 (0.92)
Confidence (no fitting)	0.70	0.84	0.67	0.81
Pairwise Dice (no fitting)	0.90	0.87	0.83	0.87
Held-out studies / failures	474 / 17	474 / 22	474 / 30	474 / 31

A conformal threshold for the check-again flag. For each model we split the test patients at random into two halves (1000 times). On one half, the threshold on the combined score (transmission view) was set as the $\lfloor \alpha(n+1) \rfloor$-th smallest score among the n failures (Dice lost ≥ 0.10), so that at most a fraction α of failures fall below it; on the other half we counted missed failures and flagged studies. With about 15 patients and their failures per half, the miss rate varies widely between splits.

Table S12: Split-conformal threshold on the combined score: failures missed (mean over splits, with the 5th–95th percentile range), and studies flagged, for target miss rates of 10% and 20%.

Target miss rate	Model	Failures missed (%)	90% range	Studies flagged (%)	Non-failures flagged (%)
10	U-Net	10	2–21	22	16
	Ensemble	9	2–19	23	17
	MC dropout	10	2–24	25	16
	nnU-Net	10	1–24	24	16
20	U-Net	20	5–36	17	12
	Ensemble	19	9–32	18	13
	MC dropout	20	6–37	20	12
	nnU-Net	20	6–40	16	9

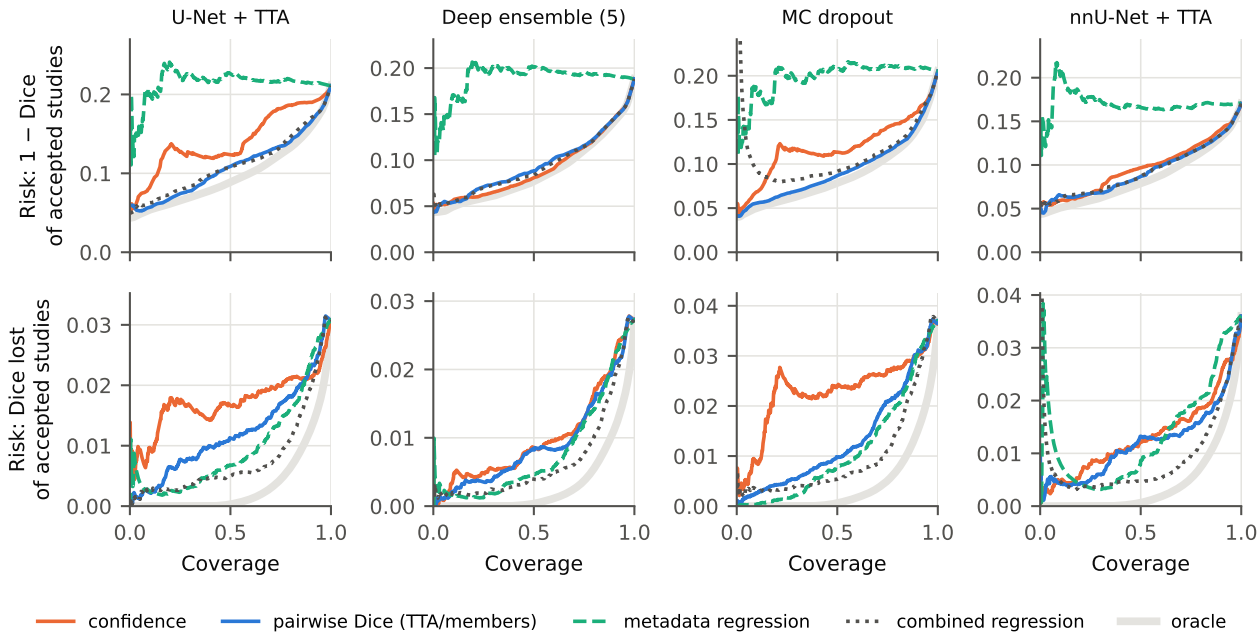


Figure S4: Risk–coverage curves for all four models. Top: absolute risk (1−Dice); bottom: transmission risk (Dice lost). Grey: oracle ranking.