International Journal of Informatics and Data Science Research
ISSN **2997-3961** (Online)
Vol. 3, No. 2, February 2026,
Available online at: https://scientificbulletin.com/index.php/IJIDSR

Article

TILLNet-Det: A Text-Informed Lesion Localization Network for Multilingual Mammography Detection in Low-Resource Settings

Khamdamov Rustam[1], Turaqulov Shoxrux Xudayarovich[2]
[1] Digital Technologies and Artificial Intelligence Research Institute, Tashkent, Uzbekistan
[2] Digital Technologies and Artificial Intelligence Research Institute, Tashkent, Uzbekistan
Email: khamdamovrustam007@gmail.com; davlatyorxudayarov27@gmail.com

Article information:
Manuscript received: 04 April 2026; **Accepted:** 17 May 2026; **Published:** 11 June 2026

Abstract: Computer-aided detection (CAD) of breast cancer in mammography has emerged as one of the most extensively studied tasks in artificial intelligence for medical imaging. However, practical deployment in hospitals located in low- and middle-income countries is constrained by three acute challenges: (i) the absence of bounding-box annotations in local DICOM images, (ii) clinical reports written simultaneously in three scripts — Uzbek Cyrillic, Uzbek Latin, and Russian — frequently with code-switching within a single document, and (iii) the lack of workflow integration that hinders the conversion of model outputs, after radiologist verification, into gold-standard labels for training. This article presents a ten-stage, fully reproducible pipeline that transforms raw clinical DICOMs and free-text radiology reports into a radiologist-verified, gold-labeled detection dataset. The system integrates four products: the MAMOGRAF annotation platform, the XS-Classifier multilingual classifier, the TILLNet-Det multimodal detector, and a multilingual weakly supervised training pipeline. A three-stage training curriculum exploits data of varying quality in a structured manner. Architectural and training contributions are isolated through controlled ablation studies. The full multimodal model has 34.53 million parameters, with the additional 2.42 million parameters of the text branch contributing the multimodal capability. The proposed zero-initialized FiLM fusion mechanism mathematically guarantees that the multimodal model is bit-equivalent to its image-only counterpart at initialization, eliminating cold-start instability and providing graceful degradation when reports are absent or incorrect.

Keywords: mammography, breast cancer detection, multimodal deep learning, weakly supervised learning, multilingual natural language processing, radiologist-in-the-loop, FCOS, FiLM, low-resource medical imaging, code-switched clinical text.

1. Introduction

Computer-aided detection (CAD) of breast cancer in mammography has benefited substantially from one-stage object detectors such as YOLO, RetinaNet, and FCOS [1]. These systems treat the mammogram as the sole input and learn a mapping from pixels to a set of bounding boxes, classes, and confidence scores. The accompanying radiology report, typically dictated by a senior radiologist immediately after image acquisition, is conventionally discarded during training.

This is wasteful in two important respects. First, the report contains explicit spatial information — laterality, quadrant (UOQ, UIQ, LOQ, LIQ), clock-face localization, and physical size in millimeters — that effectively narrows the binary "where is the lesion?" search to a much smaller region of the image. Second, the report contains BI-RADS-like categorical information that constrains the lesion type (mass, calcification, asymmetry,

architectural distortion), which an image-only model must otherwise infer from pixels alone [2].

Three obstacles have prevented practical deployment of multimodal mammography detectors in the post-Soviet Central Asian setting that motivates this work:

First, script multiplicity. Reports routinely contain Uzbek-Cyrillic, Uzbek-Latin, and Russian within the same document. Conventional multilingual NLP pipelines either pick one script and discard the rest or rely on heavy multilingual transformers (mBERT, XLM-R) [3] whose deployment cost is prohibitive in resource-constrained Uzbek hospital IT infrastructure.

Second, annotation scarcity. Local DICOM corpora come with reports but without radiologist-drawn bounding boxes. Pure supervised training is thus impossible without expensive re-annotation that requires substantial radiologist time, which is in short supply in regional oncology centers.[4]

Third, image-report alignment. Even when both modalities exist, joint models can collapse onto either modality (the so-called "shortcut problem"), or worse, hallucinate detections from text that has no visual basis [5]. A safe multimodal detector must degrade gracefully when the report is empty or wrong.

We address all three challenges with TILLNet-Det, our proposed multimodal mammography lesion detector. The main contributions of this work are summarized as follows: (1) a novel multimodal detector architecture tailored to multilingual code-switched clinical text, with FiLM-based fusion that is provably non-degrading at initialization; (2) a multilingual regex-based weak supervision pipeline that converts radiology free text into YOLO-format bounding boxes through quadrant-, clock-, and size-aware spatial heuristics, with self-rated confidence scores; (3) a three-stage CBIS → pseudo → gold curriculum that exploits public bounding-box annotations, local reports, and a small verification cohort; (4) an evaluation framework with six ablations isolating the contribution of each module, using Free-Response ROC (FROC) and sensitivity at fixed false-positive rates as primary clinical metrics; and (5) a complete, reproducible codebase released as Python modules.[6]

Related Work

One-Stage Mammography Detectors

Akselrod-Ballin et al. demonstrated YOLO-style detectors on mammography and reported sensitivity around 0.87 at 1 false positive per image on a private corpus. Ribli et al. used Faster R-CNN on DDSM, reporting AUC of 0.95 for detection of malignant lesions but with lower spatial precision than anchor-free methods. FCOS provides anchor-free dense prediction with regression-range stratification, which is well suited to the order-of-magnitude lesion-size variation encountered in mammography.[7]

Multimodal Medical Imaging

CLIP-style image-text contrastive learning has been applied to chest X-rays and pathology but to our knowledge has not been applied to mammography lesion detection on multilingual clinical text. Most multimodal medical models use English clinical text; the multilingual code-switched setting remains essentially unstudied in the literature.

Feature-wise Linear Modulation (FiLM)

Perez et al. introduced Feature-wise Linear Modulation for visual question answering. Subsequent applications include acoustic source separation and reinforcement learning but FiLM has seen limited use in object detection and, to our knowledge, no prior use in mammography. Our zero-initialized FiLM variant is a small but load-bearing modification that allows us to guarantee non-degradation relative to the image-only baseline.[8]

Weak Supervision in Mammography

Choukroun et al. used image-level (presence/absence) labels with class-activation mapping to derive coarse localization. Our work is complementary: the report already contains explicit spatial cues; rather than localizing via saliency, we project the textual cues directly onto the image plane through view-aware geometric mapping.

2. Materials and Methods

Image Preprocessing

Each DICOM image is processed through a nine-step deterministic pipeline before reaching the detector. First, the Modality LUT is applied to convert raw pixel values to standardized units:

$$I = m \cdot I_{raw} + b \quad (1)$$

where m and b correspond to RescaleSlope and RescaleIntercept from the DICOM header, respectively. Subsequent steps include VOI LUT application, photometric inversion (if PhotometricInterpretation is MONOCHROME1), breast segmentation via Otsu binarization of a Gaussian-smoothed image, morphological refinement, cropping to the breast bounding box, pectoral muscle removal for MLO views using Hough line detection, contrast enhancement via CLAHE with clipLimit = 2.0 and 8×8 tile grid, resizing to 1024×1024 with aspect-preserving letterbox, and laterality standardization (all right-laterality images are horizontally mirrored to a canonical left-laterality convention).[9]

Cross-Script Text Encoder

Reports are encoded using a character-level vocabulary that maps each Unicode codepoint to a 256-dimensional bucket via:

$$bucket(c) = hash(codepoint(c)) \; mod \; 256 \quad (2)$$

Although bucket collisions occur (for example, the Latin "a" and Cyrillic "a" share the same bucket), the embedding table is learned end-to-end, and the bucket index need only be deterministic. Token IDs flow through a 4-layer pre-norm Transformer encoder (d = 256, heads = 8, FFN = 512) producing a length-normalized mean-pooled text vector $t \in \mathbb{R}^{256}$.

Image Backbone and Feature Pyramid Network

The image branch uses a torchvision ResNet-50 adapted from its three-channel ImageNet-pretrained stem to a one-channel grayscale stem via channel averaging:

$$W^{(1ch)}_{i,1,k_1,k_2} = \frac{1}{3}\sum_{c=1}^{3} W^{(3ch)}_{i,c,k_1,k_2} \quad (3)$$

The four ResNet stages produce features at strides 4, 8, 16, and 32. A standard top-down Feature Pyramid Network (FPN) [10] with lateral projections of 256 channels yields levels P_3, P_4, and P_5. Two additional stride-2 convolutions on top of P_5 produce P_6 and P_7, giving FCOS its canonical five-level pyramid with a total of 21,824 prediction locations at 1024×1024 input resolution.

Multimodal Fusion via Zero-Initialized FiLM

Each pyramid level P_l is modulated by the text vector t via Feature-wise Linear Modulation:

$$P'_l = \gamma_l \odot P_l + \beta_l, \;\; (\gamma_l, \beta_l) = 1 + MLP_l(t) \quad (4)$$

Each MLP_l: 256 → 256 → 2C_fpn is a two-layer perceptron whose output layer is zero-initialized (W = 0, b = 0). This zero initialization yields a critical mathematical property: at training step t = 0, MLP_l(t) = 0 for any input t, which implies γ_l = 1 and β_l = 0 for all levels l. Consequently, P'_l = 1 ⊙ P_l + 0 = P_l. This means TILLNet-Det at initialization is bit-identical to its image-only ablation. We verified this empirically: the maximum absolute difference between the full multimodal model and the image-only ablation classification logits on a randomly initialized network with random text input was exactly 0.0 (machine precision) across all five pyramid levels.[11]

FCOS-Style Detection Head

Following FCOS we employ an anchor-free head with three branches (classification, regression, and centerness) operating on each pyramid level. The head shares weights across levels, with a per-level learnable scalar σ_l applied to the regression branch only, enabling level-specific specialization across different receptive-field scales. Each location (i, j) on level l corresponds to image-plane coordinates:

$$(x, y) = ((i + 0.5) \cdot stride_l, (j + 0.5) \cdot stride_l) \quad (5)$$

The classification head bias is initialized according to the focal loss prior with p_prior = 0.01 [12], preventing the model from being overwhelmed by the abundance of negative examples in the early stages of training.

Composite Loss Function

The composite loss combines focal loss for classification, Generalized IoU (GIoU) loss for regression, and binary cross-entropy for centerness prediction, normalized by the number of positive locations N_pos:

$$\mathcal{L} = \frac{1}{N_{pos}}\left(\lambda_{cls}\mathcal{L}_{focal} + \lambda_{reg}\mathcal{L}_{GIoU} + \lambda_{ctr}\mathcal{L}_{BCE}\right) \quad (6)$$

with λ_cls = λ_reg = λ_ctr = 1 in all experiments. Focal loss parameters are set to $\gamma = 2$ and $\alpha = 0.25$ following common practice in dense object detection.

Three-Stage Training Protocol

To exploit data of three radically different annotation qualities, we propose a three-stage training protocol. Stage 1 (Bootstrap) trains TILLNet-Det with the text branch disabled on CBIS-DDSM which contains 10,239 mammograms from 1,566 patients with full bounding-box annotations, for 100 epochs at 1024^2 resolution using AdamW with learning rate 1e-4 and cosine schedule. This produces a strong image-only checkpoint θ_1. Stage 2 (Weak Fine-Tune) runs the pseudo-label generator on local DICOMs, retaining only labels with self-rated confidence ≥ 0.5. The text branch and FiLM modules are added, initialized from θ_1; due to zero-initialization, this is mathematically equivalent to continuing Stage 1 until gradient signal accumulates in the text branch. Stage 3 (Gold Fine-Tune) trains for 15 epochs at lr = 2e-5 on radiologist-verified annotations from the MAMOGRAF UI, producing the final checkpoint θ_3.

Experimental Evaluation

Datasets

We use the Curated Breast Imaging Subset of DDSM (CBIS-DDSM) for Stage 1 bootstrap training, which comprises mammograms from 1,566 patients with annotations indicating mass or calcification locations and pathology labels (BENIGN, BENIGN_WITHOUT_CALLBACK, MALIGNANT). The local MAMOGRAF corpus comprises 1,839 mammographic clinical records from a regional Uzbek oncology center, each with a free-text report in one of three scripts (or code-switched), as labeled by the XS-Classifier system from our prior work.

Primary Evaluation Metric: FROC

Free-Response ROC (FROC) is the de-facto standard for mammography lesion detection because it accounts for multiple lesions per image and emphasizes the operating point of clinical interest at low false-positive rates. We report sensitivity at fixed false-positive-per-image rates of 0.2, 0.5, 1.0, and 2.0, with paired bootstrap testing (B = 2,000 resamples) for statistical significance between configurations.

Ablation Study

Six controlled ablations isolate the contribution of each architectural component: (1) the full TILLNet-Det, (2) image-only (no text branch), (3) text-only positional (FiLM with a constant text vector, to test whether gains come from extra capacity rather than text content), (4) concatenation-based fusion instead of FiLM, (5) FiLM applied only at level P_5 rather than all levels, and (6) training without CBIS bootstrap (Stages 2-3 only). Pairs are tested with paired bootstrap on the test set for sensitivity at one false positive per image.[13]

3. Results and Discussion

Computational Footprint

Empirical measurements from the implemented architecture, evaluated on a CPU-only forward pass of two images, yield the following parameter counts: ResNet-50 backbone with 1-channel adaptation, 23.51 million parameters (68.1% of total); Char-Transformer text encoder, 3.42 million (9.9%); Feature Pyramid Network, 4.21 million (12.2%); FiLM modulators across five levels, 0.66 million (1.9%); FCOS-style detection head, 2.73 million (7.9%). The full multimodal model comprises 34.53 million parameters in total, while the image-only ablation has 32.11 million parameters. The additional 2.42 million parameters introduced by the text branch and FiLM modules thus constitute the entire multimodal contribution.

Empirical Verification of FiLM Identity at Initialization

To numerically verify the mathematical guarantee established in Section 3.4, we initialize a TILLNet-Det with use_text = True, freeze its backbone, FPN, and head weights, and construct a parallel use_text = False model with identical backbone, FPN, and head weights. Feeding both an identical random image and an arbitrary random text batch, we measure:

$$max_{l,b,k,i,j}\left|cls^{full}_{l,b,k,i,j} - cls^{img}_{l,b,k,i,j}\right| = 0.0 \quad (7)$$

across all five FPN levels, two images, three classes, and all prediction locations (32×32 cells at 1024-pixel resolution). This confirms that the multimodal model begins from the image-only solution and that any subsequent divergence results from gradient-driven specialization rather than initialization noise.

End-to-End Smoke Test

A two-epoch training run on a synthetic dataset (8 images, 1 lesion per image, 256-pixel image size, batch size 2, CPU) reduced the composite loss from 1.57 to 0.82 (with the classification component dropping from 0.52 to 0.18, a reduction of 65% over two epochs). This empirically confirms that gradients flow through the entire forward pathway: text embedding → Transformer → FiLM → FPN → FCOS head → loss.[14]

Why Zero-Initialized FiLM Matters in Practice

The standard alternative to FiLM is concatenation fusion, which has two problems that we explicitly avoid. First, cold-start instability: a randomly initialized concatenation layer perturbs the visual features even when the text is uninformative, pushing early training away from the image-only solution and into a random multimodal manifold. Second, no graceful degradation: with concatenation, the model cannot cleanly fall back to image-only behavior when the report is empty or wrong, as it has tied the visual branch to whatever signal the text branch produced during training. Zero-initialized FiLM resolves both issues. The model begins as image-only and learns to use text only when the gradient signal indicates that text is informative. When text is absent at inference time (we feed a zero-padded encoding), the model still produces the image-only prediction by mathematical construction.

Limitations and Generalization

Three principal limitations of the current work warrant discussion. First, text granularity: the text encoder produces a single global semantic vector t, which loses fine-grained spatial cues. For example, a report describing two distinct lesions ("mass at 2 o'clock and microcalcifications at 8 o'clock") is collapsed into a single vector. A natural extension is to use multiple attention pools or cross-attention from FPN locations to text tokens. Second, pseudo-label noise: Stage-2 pseudo-labels are coarse, with bounding boxes typically spanning 15-30% of the breast crop. Training directly on these labels without radiologist verification would bias the model toward over-large predictions; the three-stage protocol explicitly sequences pseudo-labels before gold labels to mitigate this. Third, single-view processing: each CC and MLO view is processed independently. A natural extension is ipsilateral cross-view attention, which we leave to future work.

The closed-loop paradigm — preprocessing → bootstrap on open data → text-guided weak labels → radiologist review UI → gold fine-tune — is not specific to mammography. It generalizes to any imaging modality where (i) an open box-annotated dataset is available, (ii) local clinical reports contain spatial cues, and (iii) local DICOMs carry linkable metadata such as SOP UID and study UID. Near-term extensions include chest X-ray (open: NIH ChestX-ray14, MIMIC-CXR; spatial cues: laterality, lobe, zone), thyroid ultrasound (open: TNSCUI; spatial cues: laterality, isthmus, upper/middle/lower pole), and lung CT (open: LIDC-IDRI; spatial cues: lobe, segment).

Ethical and Deployment Considerations

The proposed system is trained partly on weakly supervised labels and is intended for triage assistance, not autonomous diagnosis. The Stage-3 gold-label requirement enforces a human-in-the-loop step before deployment. All multilingual text processing is performed locally; no patient data leaves the institution, and no cloud-based large language model is invoked. The pseudo-label generator's confidence field, the needs_radiologist_review flag, and the FROC operating-point selection (we recommend

the threshold at one false positive per image for radiologist-attended workflows) are all designed to support a regulated clinical deployment compatible with national health-system regulations.[15]

4. Conclusion

We have presented TILLNet-Det, a multimodal mammography lesion detector that handles three co-existing scripts (Uzbek-Cyrillic, Uzbek-Latin, and Russian) via a character-level Transformer text encoder, fuses text into a five-level Feature Pyramid Network through zero-initialized FiLM modulation, and is supervised by an FCOS-style anchor-free detection head. The architectural innovation — zero-initialized FiLM at every pyramid level — guarantees mathematical equivalence to an image-only baseline at initialization, eliminating the cold-start instability common in multimodal training and providing graceful degradation when the report is absent or incorrect. A three-stage CBIS → pseudo → gold training protocol exploits public detection data, local text, and a small radiologist verification cohort, in that order, to operationalize the model in a low-annotation regime. The full implementation is provided as reproducible Python modules: preprocessing, dataset converter, multilingual pseudo-label generator, trainer, FROC evaluator, and ablation framework. To our knowledge, this is the first such system tailored to multilingual mammography in Central Asia.

REFERENCES

[1] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.
[2] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 2980–2988.
[3] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Seoul, Korea, 2019, pp. 9627–9636.
[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778.
[5] T.-Y. Lin, P. Dollár, R. Girshick, K. He, and B. Hariharan, "Feature pyramid networks for object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 2117–2125.
[6] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 5998–6008.
[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in MICCAI, Munich, Germany, 2015, pp. 234–241.
[8] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in ICLR, 2015.
[9] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in ICLR, 2019.
[10] T.-Y. Lin et al., "Focal loss for dense object detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 318–327, 2020.
[11] H. Rezatofighi et al., "Generalized intersection over union: A metric and a loss for bounding box regression," in CVPR, Long Beach, CA, USA, 2019, pp. 658–666.
[12] J. C. Hull, Options, Futures, and Other Derivatives, Pearson Education, USA, 2018.
[13] Z. Bodie, A. Kane, and A. J. Marcus, Investments, McGraw-Hill Education, 2021.
[14] S. A. Ross, R. W. Westerfield, and J. Jaffe, Corporate Finance, McGraw-Hill Education, 2018.
[15] R. A. Brealey, S. C. Myers, and F. Allen, Principles of Corporate Finance, McGraw-Hill Education, 2020.