AOI: 10.100.234513.0253
13th international conference on Microelectronics, Circuits and Systems, Micro2026

Reducing Diagnostic Errors in Medical Imaging via Selective Prediction

Satyendra Yadav, Vidushi Sharma, Rajiv Ratn Shah

Department of Computer Science & Engineering
Gautam Buddha University, Greater Noida, INDIA
Corresponding Author: vidushi@gbu.ac.in

ABSTRACT

Deep learning models have achieved remarkable success in medical image analysis in recent times. However, their tendency to produce overconfident predictions even under high uncertainty, limits their reliability in safety-critical applications. In the clinical setting, minimizing the incorrect predictions is often more important than maximizing the overall coverage. This paper investigates the use of selective prediction via training dynamics (SPTD) to improve diagnostic reliability by enabling models to abstain from uncertain predictions. Unlike the conventional confidence based approaches, SPTD leverages prediction instability across training checkpoints as a proxy for uncertainty. Experiments are conducted on three different medical imaging datasets including diabetic retinopathy, urine cell classification and bone tumor classification datasets. Results clearly demonstrate that SPTD has consistently improve the prediction accuracy at lower coverage levels and achieved more than 20% improvement compared to non-SPTD model. The findings highlight the effectiveness of training dynamics as a robust uncertainty estimation mechanism and establish selective prediction as a viable strategy for reducing diagnostic errors in healthcare AI systems.

Keywords: Deep learning, Medical Imaging, Selective Prediction, Training Dynamics, Uncertainty Estimation, Abstention.

I. INTRODUCTION

Recent advances in deep learning have significantly improved performance in medical image analysis tasks such as disease classification, lesion detection and organ segmentation etc. Even with these advancements, modern neural networks many a times exhibit overconfidence and they produce predictions even when making predictions for ambiguous inputs on which they are not even certain. In healthcare applications, such behavior can lead to severe situations like misdiagnosis and inappropriate treatment decisions. This is highly worrisome in high-stake situations like healthcare where such errors can cause significant risk to the patient safety and can be even life threatening.

To address this limitation, many researchers have explored alternatives that make it possible for the decision makers to make their own decisions in the cases where AI predictions can go wrong. One such popular approach is selective prediction [1], which introduces a reject option that allows models to abstain from making predictions when uncertainty is high. In the standard AI workflow, the predictions are made for every input but in case of selective prediction there are mechanisms to identify if the prediction is reliable or not and basis this information it abstains and directs the human to make a decision on their own.

The Selective prediction method is depicted in Figure. 1, where the model decides to predict or abstain based on the calculated uncertainty. This paradigm shifts the objective from maximizing accuracy on all samples to optimizing the risk–coverage trade-off, where predictions are made only when they are reliable.

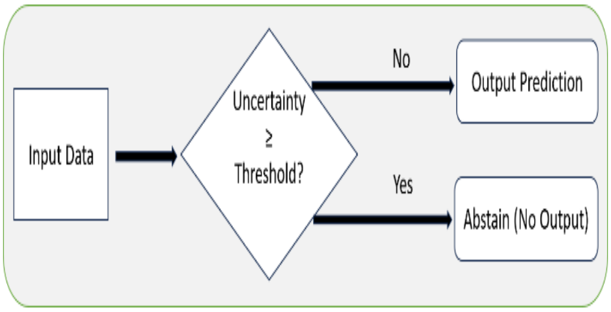

Fig. 1. Selective prediction method

In this work, we have used and evaluated Selective Prediction via Training Dynamics (SPTD) [2], a recently proposed framework that estimates uncertainty based on prediction consistency during the training phase. In contrast to the traditional approaches that rely on softmax confidence scores, SPTD captures temporal instability in predictions during the model training, and then it is correlated with sample difficulty and the risk of misclassification of the input sample. It is an uncertainty estimation method that is helpful in identifying unreliable machine predictions by doing analysis of prediction disagreement across checkpoints during the training. Hence, SPTD utilizes the training dynamics which denotes how a model has evolved during the training.

The key contributions of this paper are as follows:

- We have applied the SPTD method to three different medical imaging datasets spanning different modalities and disease.
- We have performed a comprehensive evaluation of the SPTD method across varying coverage levels to analyses the risk–coverage trade-off.

ISBN: 978-81-689089-0-1

- We have demonstrated that the SPTD method significantly reduces diagnostic errors by filtering unreliable predictions.

II. RELATED WORK

The concept of classification with the abstain option was first introduced by Chow et al. [3], and it has established the strong theoretical foundation for selective prediction. Later works have extended this foundation to deep learning including Selective Net [4] and learning-to-defer approaches, which integrate abstention into the model training. Uncertainty estimation has been widely studied in medical AI using techniques such as Bayesian neural networks, Monte Carlo dropout, and deep ensembles. While these methods provide uncertainty estimates, they often require additional computational overhead and are sensitive to calibration issues. Researchers have also explored ways like data augmentation [5], scaling [6] and pre-processing to improve the performance but they fail to achieve gains beyond a certain point.

Researchers have acknowledged the potential [7] of selective prediction systems that defers the decision making to the human expert when AI predictions are unreliable. Bondi et al. [8], have explained the importance of considering how to communicate the decision to abstain while designing the selection prediction and related systems. Further, Jabbour et al. [9] have compared the human expert performance without involving AI, first with the AI assisted performance and then with AI assisted with selective prediction. Their results have shown that the clinicians underdiagnose when they are informed that the AI has abstained this prediction. Wang et al. [9] have also supported selective prediction to improve the AI-assisted performance with human intervention. They have first collected the decisions made by the humans and AI separately and then calculated the AI-assisted performance by simulating different human-AI teaming strategies.

Kye et al. [11] have shown that the training dynamics contain valuable information about data difficulty and model confidence. SPTD leverages this insight by analyzing prediction consistency across epochs, offering a model-agnostic and calibration-free alternative to traditional uncertainty estimation methods. SPTD analyses how consistently a model predicts the given input in different training epochs. The core idea here is that if the predictions are stable and does not change epoch by epoch then they are easy samples and can be predicted correctly by the AI system. On the other hand, in case the predictions are changing frequently then it shows that this input is difficult and the AI system is uncertain with its predictions. In order to quantify these findings SPTD calculates the disagreement score by comparing the predictions made at different checkpoints and the final output. The disagreements which occur in later epochs get higher importance because even after a good amount of training still the model is unable to make stable predictions. This disagreement score serves as the measure of uncertainty in the input and basis the threshold the model decides to predict or to abstain.

The objective of this research is to utilize the capabilities of SPTD and apply it on different medical imaging datasets to evaluate its effectiveness and performance in medical imaging. We will also analyses the risk-coverage trade-off to see the performance we can achieve with different coverage levels. Further, we will see if we are able to reduce the diagnostic errors and help in filtering out the unreliable or wrong predictions.

III. METHODOLOGY

In this section, we will start with explaining the step-by-step method of selective prediction in healthcare. Further, this section also provides the information about the dataset used in this work and the deep learning model used for evaluation on these different datasets. In this research, we have used SPTD to reduce the prediction error in the deep learning models to enhance their accuracy so that they can choose not to predict when they are uncertain and to predict only when they are confident enough in the prediction they are making.

A. Selective prediction

In Selective prediction the AI model is allowed to make predictions when it is confident and abstain when it is uncertain. By allowing the AI model to abstain, selective prediction achieves better diagnostic reliability.

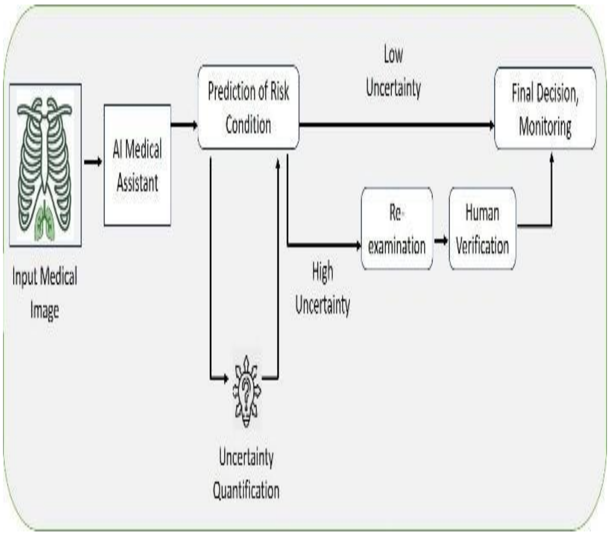

Fig. 2. Step by step Selective prediction method in healthcare

Figure. 2 depicts the end-to-end selective prediction method. The input medical image is provided as the input to the AI model and it first calculates the uncertainty based on different parameters and then based on the uncertainty score it moves to the next step. If the uncertainty is below the threshold, it allows the AI model to predict and share its results. However, if the

ISBN: 978-81-689089-0-1

uncertainty is above the threshold limits, then it abstains and suggests re-examination with the human expert.

TABLE I
CLASS WISE TRAIN, VALIDATION AND TEST SPLIT OF ALL THREE DATASETS

Class	Train	Test	Valid	Total
APTOS diabetic retinopathy dataset				
No DR	1430	277	91	1798
Mild	276	46	19	341
Moderate	758	136	40	934
Severe	143	30	9	182
Proliferative DR	218	42	19	279
Total	2825	531	178	3534
Urine cell classification dataset				
HGUC	286	65	37	388
Keratin	27	5	5	37
Leukocyte	20	5	1	26
Normal	457	141	78	676
Squamous	47	16	5	68
Suspicious	482	144	63	689
Total	1319	376	189	1884
ORS Bone Tumor Classifier dataset				
MALIGNANT-osteosarcoma	550	70	31	651
MALIGNANT-chordoma	469	43	27	539
BENIGN-fibrous dysplasia	428	47	27	502
BENIGN-nonossifyingfibroma	375	47	21	443
BENIGN-hemangioma	369	36	28	433
BENIGN-giantcelltumor	342	48	25	415
BENIGN-chondroblastoma	331	47	24	402
HEALTHY-bonecompact	322	47	15	384
MALIGNANT-ewingsarcoma	310	38	16	364
MALIGNANT-multiple myeloma	300	31	20	351
MALIGNANT-chondrosarcoma	292	27	18	337
MALIGNANT-adamantinoma	280	33	14	327
BENIGN-enchondroma	237	25	6	268
BENIGN-aneurysmalbonecyst	211	21	11	243
BENIGN-unicameralbonecyst	158	25	10	193
BENIGN-aneurysmalbonecyst BENIGN-fibrous dysplasia	33	4	1	38
BENIGN-pagetsdisease	9	2	0	11
Total	5016	591	294	5901

A. *Dataset description*

In this research, we have used three open-source medical imaging datasets for different disease classification. The first dataset is APTOS diabetic retinopathy dataset [12] which contains a total of 3534 images for five different classes of diabetic retinopathy. The second dataset is Urine cell classification dataset [13] and it contains a total of 1884 images from six different classes. This is the smallest dataset in number of images but has an additional class as compared to APTOS dataset. The third dataset is the ORS Bone Tumor Classifier dataset [13] and this is the largest dataset in terms of both the number of images and number of classes in it. It contains a total of 5901 images of seventeen different classes. Table I shows the dataset used with their classes and count of images in each class. It also provides the exact number of Training, Validation and Test subset.

B. *Models training and evaluation*

In this study, we have used YOLOv11, a state-of-the-art object detection model from YOLO family models. We have combined it with SPTD to improve the disease detection and classification in healthcare AI. We have tried to compare both the results to showcase the effectiveness of SPTD approach over YOLOv11 being used independently. We have trained both YOLOv11 and YOLOv11 with SPTD approach on the all three datasets in the same conditions for fair and valid performance comparison. We have used a batch size of 256 for APTOS and Urine datasets, but batch size of 512 is used with the ORS bone dataset as it was the large dataset. However, the models were trained for 50 epochs on all the three datasets.

IV. RESULTS AND ANALYSIS

This section offers a detailed review of the performance of YOLOv11 and YOLOv11 with STPD on three publicly available medical imaging datasets. The results were evaluated using standard object detection metrics accuracy as in selective prediction it is important to understand the accuracy of the predictions made by the model. If it’s uncertain then it must abstain to keep the accuracy high. In the healthcare domain it matters the most as the wrong predictions in the worst case can be life threatening.

A. *Performance evaluation on APTOS diabetic retinopathy dataset*

In our coverage. On the other hand, YOLOv11 without SPTD was at just above 75% accuracy. Hence it is experiments; YOLOv11 with selective prediction was able to achieve its best accuracy more than 95% with 50% clear that we are able to improve accuracy by more than 25% by reducing the coverage to half. Table II depicts the overall accuracy achieved on different coverage levels. With the reduced coverage we can observe a clear increase in the accuracy.

ISBN: 978-81-689089-0-1

TABLE II
ACCURACY COMPARISON ON APTOS DATASET FOR VARIED COVERAGE

Coverage (%)	Accuracy		Accuracy Gain (%)
	YOLOv11	YOLOv11 with SPTD	
50	75.57	95.45	26.31
70	75.57	85.37	12.97
80	75.57	80.85	6.99
90	75.57	79.11	4.68
100	75.57	76.14	0.75

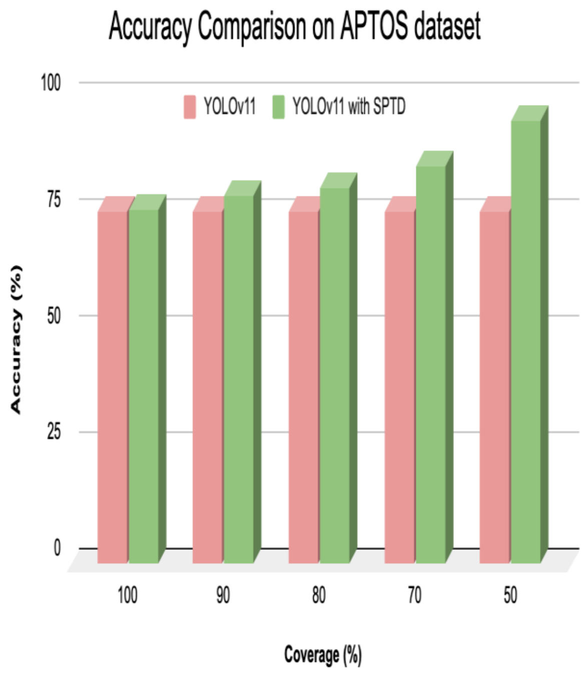

Fig. 3. Accuracy comparison on APTOS diabetic retinopathy dataset for varied coverage.

Comparison in terms of accuracy is shown in Figure. 3. We can observe that how YOLOv11 with SPTD is performing well as compared to YOLOv11 without SPTD. The increase in accuracy is also consistent and even getting better with the decrease in the coverage levels.

B. Performance evaluation on urine cell classification dataset

In the Urine dataset which is the smallest among all three dataset we are not seeing huge performance improvements as we were seeing in the APTOS dataset. In this case, we have 77.89% of top accuracy for YOLOv11 with SPTD, whereas for YOLOv11 without SPTD is just above 65% with the coverage level of 50%. This is around 20% gain for SPTD as compared to the non-SPTD variant; Table III mentions all these above mentioned results. Figure. 4 depict the accuracy comparison graph for the Urine dataset which also confirms that the gains are achieved with the SPTD variant, but these gains are not substantial.

TABLE III
ACCURACY COMPARISON ON URINE DATASET FOR VARIED COVERAGE

Coverage (%)	Accuracy		Accuracy Gain (%)
	YOLOv11	YOLOv11 with SPTD	
50	65.08	77.89	19.68
70	65.08	71.21	9.42
80	65.08	69.54	6.85
90	65.08	68.82	5.75
100	65.08	64.55	-0.81

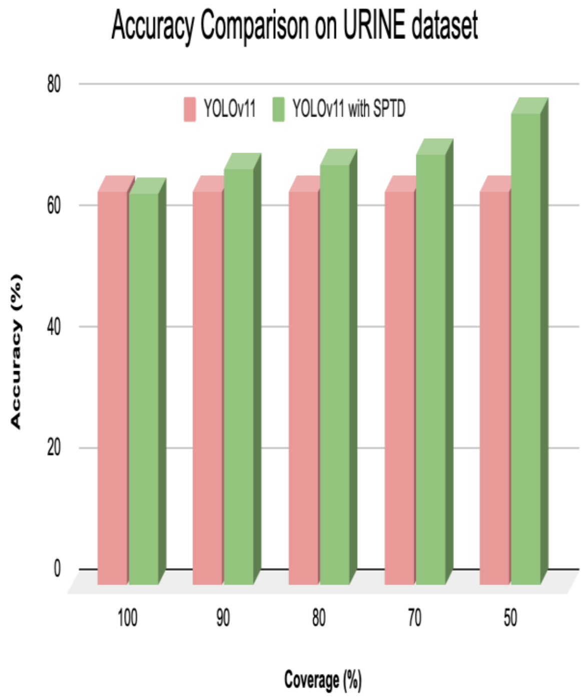

Fig. 4. Accuracy comparison on Urine cell classification dataset for varied coverage.

C. Performance evaluation on ORS Bone Tumor Classifier dataset

Table IV summarizes the accuracy results for the largest ORS dataset but the accuracy gains with SPTD are minimum here. We can observe the best accuracy as 88.96% in the case of 50% coverage and it is around 10% gain as compared to the non-SPTD variant. In all the coverage levels performance gain can be observed clearly, however not much difference among 80% and 90% coverage levels. This dataset was largest but the number of classes are also 17 and some of the classes are very less represented in the dataset. Some of the Benign classes are even less than 1% of the total images and due to this the accuracy gains with SPTD are not so strong. In our error analysis we found that most of the wrong predictions belong to the less represented classes in this dataset. Figure. 5 show the accuracy comparison on the ORS bone tumour classifier dataset.

ISBN: 978-81-689089-0-1

TABLE IV
ACCURACY COMPARISON ON ORS DATASET FOR VARIED COVERAGE

Coverage (%)	Accuracy		Accuracy Gain (%)
	YOLOv11	YOLOv11 with SPTD	
50	80.34	88.96	10.73
70	80.34	85.51	6.44
80	80.34	83.05	3.37
90	80.34	83.02	3.34
100	80.34	79.66	-0.85

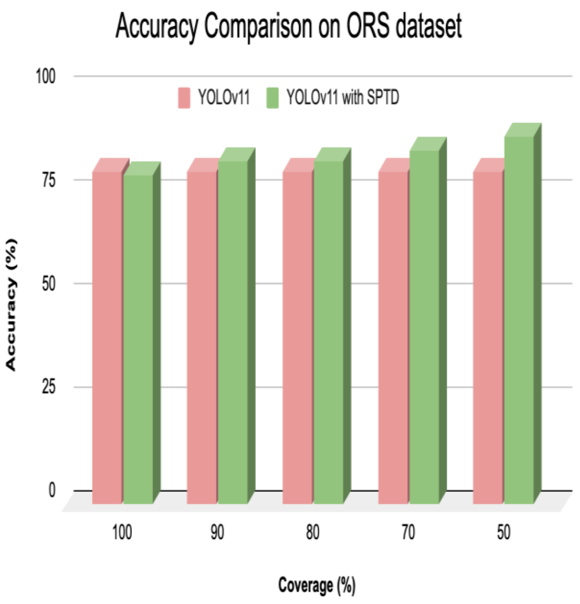

Fig. 5. Accuracy comparison on ORS bone tumor classifier dataset for varied coverage.

V. CONCLUSION AND FUTURE WORK

This paper demonstrates that selective prediction via training dynamics significantly enhances the reliability of medical imaging systems. By allowing the model to abstain from uncertain predictions, SPTD reduces diagnostic errors and improves accuracy at lower coverage levels. The performance evaluation on multiple datasets confirms the effectiveness of SPTD as a robust uncertainty estimation technique. This approach is very much valuable in safety-critical applications such as healthcare, where incorrect predictions can be fatal and life threatening.

In the future, we can evaluate performance on some large and complex diverse datasets. Another possible future work can be extending the SPTD method to object detection and segmentation tasks. We can also work on improving the uncertainty estimations to move it closer to near perfect results. Another aspect can be identifying the right coverage to achieve 100% accuracy, which will ensure using AI techniques in real-world clinical workflows.

ACKNOWLEDGEMENT

This document is prepared in support of Applied Computer Technology, a research oriented technology based company. All available DOI/AOI numbers should be included within references.

REFERENCES

[1] El-Yaniv, R.; and Wiener, Y. 2010. On the Foundations of Noise-free Selective Classification. Journal of Machine Learning Research, 11(5).

[2] Rabanser S, Thudi A, Hamidieh K, Dziedzic A, Bahceci I, Sediq AB, Sokun H, Papernot N. Selective Prediction via Training Dynamics. Transactions on Machine Learning Research. 2025.

[3] Chow, C. K., "On optimum recognition error and reject tradeoff," IRE Transactions on Information Theory, vol. 16, no. 1, pp. 41–46, Jan. 1970.

[4] Geifman Y, El-Yaniv R. Selectivenet: A deep neural network with an integrated reject option. In International conference on machine learning 2019 May 24 (pp. 2151-2159). PMLR.

[5] Yadav S, Saini I, Sharma V, Shah RR. Evaluating YOLOv12 and RT-DETR for Robust Brain Tumor Detection in MRI Scans. In2026 2nd International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI) 2026 Feb 12 (pp. 1085-1090). IEEE.

[6] Yadav S, Saini I, Sharma V, Shah RR. Boosting Automated Urine Sediment Classification via Data Scaling. In2025 Second International Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT) 2025 Dec 4 (pp. 374-379). IEEE.

[7] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems. 2017;30.

[8] Bondi E, Koster R, Sheahan H, Chadwick M, Bachrach Y, Cemgil T, Paquet U, Dvijotham K. Role of human-AI interaction in selective prediction. In Proceedings of the AAAI Conference on Artificial Intelligence 2022 Jun 28 (Vol. 36, No. 5, pp. 5286-5294).

[9] Jabbour S, Fouhey D, Banovic N, Shepard SD, Kazerooni E, Sjoding MW, Wiens J. On the limits of selective AI prediction: a case study in clinical decision making. arXiv preprint arXiv:2508.07617. 2025 Aug 11.

[10] Wang DY, Ding J, Sun AL, Liu SG, Jiang D, Li N, Yu JK. Artificial intelligence suppression as a strategy to mitigate artificial intelligence automation bias. Journal of the American Medical Informatics Association. 2023 Oct 1;30(10):1684-92.

[11] Kye SM, Choi K, Byun H, Chang B. TiDAL: Learning training dynamics for active learning. In Proceedings of the IEEE/CVF international conference on computer vision 2023 (pp. 22335-22345).

ISBN: 978-81-689089-0-1

[12] Karthik, Maggie & Dane, S. Aptos 2019 blindness detection. https://kaggle.com/competitions/aptos2019-blindness-detection (2019). Kaggle.

[13] "Urine Cell Classification Computer Vision Dataset," https://universe.roboflow.com/eyes-classification/urine-cell-classification-prgqm-u6ryy , December 2025, visited on 2026-04-18. [Online]. Available: https://universe.roboflow.com/eyes-classification/urine-cell-classification-prgqm-u6ryy

[14] "ORS-Bone-Tumor-Classifier Computer Vision Dataset," https://universe.roboflow.com/classification-tests/ors-bone-tumor-classifier-gsbfu , December 2025, visited on 2026-04-18. [Online]. Available: https://universe.roboflow.com/classification-tests/ors-bone-tumor-classifier-gsbfu

Cite this article as:

Satyendra, Vidushi *and et. al. "Reducing Diagnostic Errors in Medical Imaging via Selective Prediction", Proceedings of* 13th international conference on Microelectronics, Circuits and Systems, Micro2026

Displayed as online on 15th July 2026.

Link: http://actsoft.org/science/act2026-pro/511-micro2026.pdf

ISBN: 978-81-689089-0-1