TASK-AWARE COLOR CONSTANCY FOR CHRONIC KIDNEY DISEASE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Smartphone-based kidney health prediction offers a low-cost alternative to laboratory testing, but its accuracy degrades significantly under unconstrained lighting conditions that distort the colorimetric signals of creatinine test strips. Standard color constancy methods, which optimize for perceptual visual quality, often fail to preserve the precise spectral cues required for clinical quantification. To address this, we propose a Task-Aware Color Constancy (TACC) framework that explicitly aligns illumination correction with downstream disease prediction. Our approach integrates a conditional generative adversarial network (TACC-GAN) with NeurSPR, a neural structured prediction module. Crucially, we introduce a differentiable histogram layer based on Kernel Density Estimation (KDE), which bridges the "non-differentiable bottleneck" of traditional regressors. This allows diagnostic losses to backpropagate directly to the generator, optimizing color correction for clinical accuracy rather than visual appeal. We evaluate on a demographically stratified dataset constructed from US population statistics. Among deployable (reference-free) methods, TACC achieves the best kidney-health classification performance in every lighting condition, raising the F1-score from 0.50 to 0.87 under extreme lighting variations while keeping creatinine-estimation error competitive with the strongest baseline. Notably, it approaches oracle calibration methods that require a white-light reference at test time, despite using no such reference, opening the door towards reliable point-of-care monitoring under real-world settings.

1 INTRODUCTION

Chronic Kidney Disease (CKD) affects approximately 14% of adults in the United States, yet nearly 90% of affected adults remain unaware of their condition until advanced stages Centers for Disease Control and Prevention (2023). Early diagnosis is therefore critical, but clinical chemistry analyzers remain costly for rural and low-resource settings, ranging from $15,000 to more than $100,000 LabX (2025).

Smartphone-based point-of-care (PoC) testing with colorimetric creatinine strips offers a low-cost alternative. These tests rely on the Jaffe reaction, where creatinine reacts with picric acid to produce an orange-red colorimetric shift. Prior work using Selective Partitioned Regression (SPR) achieved strong performance under controlled lighting Whelan et al. (2024; 2022), but real-world use remains limited by ambient illumination and smartphone ISP distortions that alter the chromatic cues required for quantification. Standard color constancy methods optimize perceptual quality rather than diagnostic fidelity, and classical diagnostic models (Random Forests, Gradient Boosted Trees) are non-differentiable, preventing diagnostic errors from being backpropagated into the correction module. We propose Task-Aware Color Constancy (TACC), which replaces these static classifiers with NeurSPR, a neural structured-prediction module whose differentiable KDE histogram layer lets diagnostic losses flow backward to the image generator, forming a closed feedback loop that optimizes preprocessing to maximize classification accuracy.

Our contributions are threefold: (1) **Task-Aware Color Constancy (TACC-GAN):** We propose a generative framework utilizing mask-guided supervision to enforce spectral fidelity within biochemically active zones, preventing hallucinations in critical diagnostic regions. (2) **End-to-End Differentiable Diagnostics:** We overcome the non-differentiable bottleneck by introducing NeurSPR,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

which employs Gaussian Kernel Density Estimation (KDE) for differentiable histogram features, enabling direct gradient flow from clinical prediction loss to the color correction module. (3) **Rigorous Evaluation on Stratified Data:** We evaluate our method on a demographically diverse, US-aligned dataset with a strict patient-level split to validate generalization under unconstrained lighting variations.

2 Related Work

2.1 On-Device Kidney Health Monitoring

Whelan et al. Whelan et al. (2024) established the baseline for smartphone-based creatinine quantification with Selective Partitioned Regression (SPR). While effective in controlled laboratory settings, SPR relies on static, non-differentiable decision trees that cannot adapt to the non-linear Image Signal Processing (ISP) distortions introduced by mobile cameras. In unconstrained environments, these ISP effects cause significant colorimetric shifts that degrade diagnostic accuracy Karaimer & Brown (2016). We advance their idea by evolving SPR into **NeurSPR**, a fully differentiable neural module. Unlike prior static approaches, NeurSPR allows diagnostic losses to backpropagate directly into the preprocessing pipeline, enabling end-to-end optimization for clinical fidelity.

2.2 Generative Color Correction and Medical Normalization

A major branch of the field focuses on *Illuminant Estimation*, where models like FC4 Hu et al. (2017) and C4 Yu et al. (2020) regress a global 1×3 light source vector. While effective for general scenes, these methods often require calibrated ground-truth (GT) illumination labels, which are unavailable in many medical datasets.

A parallel challenge exists in digital pathology, where *Stain Normalization* techniques–such as StainGAN Shaban et al. (2019)–use generative adversarial networks to correct scanner-induced color variations in H&E slides. Inspired by these domain-specific normalization techniques, our work follows the *Illuminant Correction* paradigm, mapping input images directly to a white-balanced diagnostic space. We compare against **AngularGAN** Sidorov (2019), which applies GANs for direct restoration, and **Deep White-Balance (DeepWB)** Afifi & Brown (2020), which utilizes a multi-decoder design for sRGB editing. We distinguish our approach by prioritizing the preservation of biochemical color cues over perceptual pleasantness or scene-wide illuminant estimation.

2.3 Task-Aware ISP and Differentiable Diagnostics

Unsupervised Domain Adaptation aligns feature distributions globally but may not preserve the local radiometric integrity required for the Jaffe reaction complex **?**. Our framework instead follows the Task-Aware Image Signal Processing (TA-ISP) paradigm, where image enhancement is optimized for downstream task performance rather than visual appeal Zhang et al. (2024). Similar task-driven objectives have been effective in medical reconstruction tasks such as MRI and CT segmentation Liang & Lauterbur (2000).

A key barrier in existing CKD pipelines is the non-differentiability of classical classifiers. We address this using differentiable KDE histogram layers, replacing hard binning with soft assignments that allow diagnostic gradients to flow through NeurSPR into TACC-GAN Peeples et al. (2021); Wang et al. (2016). This enables color correction to be optimized directly for CKD classification.

3 Method

We estimate creatinine concentration and kidney health status from mobile-captured strip images under uncontrolled lighting by learning a color-constancy transformation to a canonical white-light reference explicitly aligned with accurate kidney health prediction. As illustrated in Figure 1, our framework couples two components: **TACC-GAN**, a task-aware color constancy module using a conditional GAN to map images from heterogeneous lighting to a canonical white-light reference domain and **NeurSPR**, a neural structured-prediction module that processes histogram features from

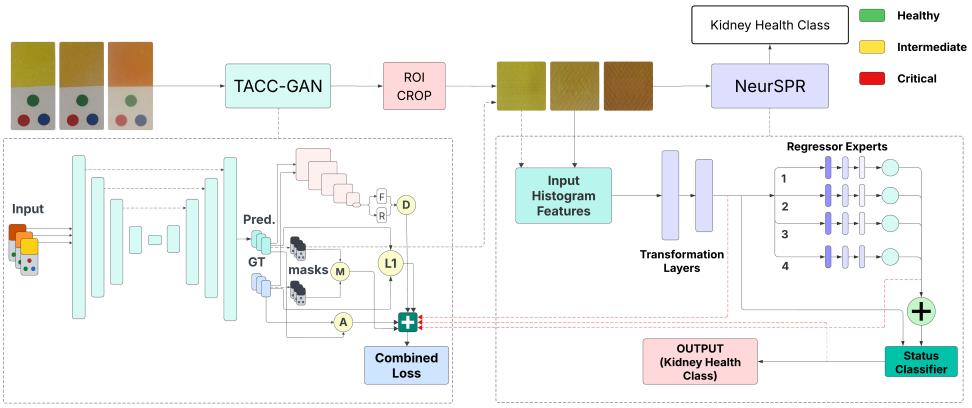


Figure 1: Overview of the proposed task-aware color constancy and CKD prediction pipeline. TACC-GAN performs canonical lighting transformation under heterogeneous lighting, followed by ROI cropping, histogram feature extraction, and concentration estimation with severity class prediction using NeurSPR. Task-aware feedback from NeurSPR guides the optimization of the color constancy module.

the transformed images to estimate creatinine via bin-aware regression and classify kidney health using established clinical thresholds.

3.1 TACC-GAN: Task-Aware Color Constancy

We implement color correction using a conditional Generative Adversarial Network (TACC-GAN) designed to map input images A (captured under unconstrained lighting) to a canonical white-light reference domain B. Unlike standard image-to-image translation models that treat all pixels equally, our generator G is spatially constrained to preserve the biochemical integrity of the test strip.

To prevent the generator from "hallucinating" colors in the critical reaction zones, we incorporate a tripartite spatial mask $M \in \{m_{det}, m_{rgb}, m_{ref}\}$. This mask isolates the Detection Zone (m_{det}), where the creatinine reaction occurs; the RGB Chromatic Anchors (m_{rgb}), used for calibration; and the White Reference Area (m_{ref}). These masks are applied during the loss calculation to enforce strict spectral fidelity in diagnostic regions while allowing more flexibility in the background.

3.2 NeurSPR: Differentiable Structured Prediction

The core innovation of our framework is **NeurSPR**, a fully differentiable evolution of the Selective Partitioned Regression (SPR) algorithm. NeurSPR replaces decision trees with a neural architecture, enabling task-aware gradients to flow from the final diagnosis back to the raw image pixels.

Differentiable Histogram via Kernel Density Estimation (KDE) Standard histograms are non-differentiable binning operations that break the gradient chain. To resolve this, we implement a Differentiable Histogram Layer using Gaussian Kernel Density Estimation (KDE).

Given a color corrected image feature map X, the value of the k-th bin in the histogram H is computed as the sum of soft-assignments of all pixels $x_i \in X$,

$$H_k = \sum_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|x_i - \mu_k\|^2}{2\sigma^2}\right), \tag{1}$$

where μ_k represents the fixed center of the k-th bin (determined by the clinical range of the Jaffe reaction) and σ is the smoothing bandwidth. This formulation ensures that H_k is a smooth function of the pixel intensities x_i, allowing gradients $\frac{\partial \mathcal{L}}{\partial x_i}$ to flow through the histogram generation process.

Mixture-of-Experts Regression NeurSPR predicts creatinine concentration using a soft Mixture-of-Experts (MoE) formulation. The model consists of a shared feature extractor and K expert regressors, where each expert E_k specializes in a specific concentration sub-interval. A gating network

computes a softmax probability distribution p over the experts. The final estimated creatinine concentration $\hat{y}$ is the probability-weighted sum of expert outputs,

$$\hat{y} = \sum_{k=1}^{K} p_k \cdot E_k(H). \tag{2}$$

This soft-weighting maintains differentiability, unlike the hard argmax selection used in classical SPR.

3.3 CKD Classification

Unlike standard classification tasks, the CKD classification is constrained by physiological parameters derived from the estimated Glomerular Filtration Rate (eGFR), which is a function of the predicted creatinine concentration and several other factors. Following the Modification of Diet in Renal Disease (MDRD) equation Levey et al. (2007), we define the eGFR as a non-linear function of the estimated serum creatinine (S_{Cr}), age, and patient demographic coefficients,

$$\begin{aligned} \text{eGFR} = 175 \times (S_{Cr})^{-1.154} \times (\text{age})^{-0.203} \\ \times [0.742 \text{ if female}] \times [1.212 \text{ if Black}] \end{aligned} \tag{3}$$

where S_{Cr} represents the serum creatinine concentration in mg/dL. To optimize the TACC-GAN generator for diagnostic utility, we define three sets of disjoint clinical status based on the calculated eGFR thresholds Whelan et al. (2024): **Critical** (eGFR ≤ 15), **Intermediate** ($15 <$ eGFR ≤ 60), and **Healthy** (eGFR > 60) mL/min/1.73m^2.

3.4 Loss Functions and Optimization Objectives

The proposed framework is optimized using a combination of image-level and task-level loss functions that jointly supervise canonical lighting transformation and downstream creatinine estimation. The total objective function combines visual stability losses with diagnostic accuracy losses to ensure the generated images are both realistic and clinically valid.

Visual Stability Losses: To ensure the generated images resemble the white-light ground truth, we employ:

1. *Adversarial Loss* ($\mathcal{L}_{GAN}$) forces the generator to produce photorealistic outputs.
2. *Masked Reconstruction Loss* ($\mathcal{L}_{L1}^{masked}$) enforces high precision in the masked diagnostic regions (M) compared to the background.
3. *Angular Consistency Loss* ($\mathcal{L}_{ang}$) penalizes angular deviations in the RGB vectors to prevent hue shifts that mimic pathological color changes.

Clinical Anchoring Losses (Task-Aware Feedback): To optimize for diagnostic utility, we backpropagate the NeurSPR prediction error:

4. *Diagnostic Loss* ($\mathcal{L}_{NeurSPR}$), composed of the regression error ($\mathcal{L}_{reg}$) and the kidney health status classification error ($\mathcal{L}_{status}$).

The total TACC-GAN objective is thus defined as,

$$\mathcal{L}_{total} = \lambda_{GAN}\mathcal{L}_{GAN} + \lambda_g\mathcal{L}_{L1}^{global} + \lambda_m\mathcal{L}_{L1}^{masked} + \lambda_a\mathcal{L}_{ang} + \lambda_n\mathcal{L}_{NeurSPR}. \tag{4}$$

In the following we provide further details on each loss.

Adversarial Loss Isola et al. (2017) TACC-GAN is trained using a conditional adversarial objective. Given an input image A captured under unknown illumination and its corresponding canonical white-light reference B, the discriminator D is trained to distinguish real pairs (A, B) from generated pairs $(A, G(A))$. The adversarial loss is defined as

$$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{A,B}[\log D(A, B)] + \mathbb{E}_A[\log(1 - D(A, G(A)))]. \tag{5}$$

Reconstruction Loss Isola et al. (2017); Zhao et al. (2016) We enforce both global color consistency and local accuracy in diagnostically relevant regions using a combined reconstruction loss:

$$\mathcal{L}_{\text{L1}}^{\text{global}} = \|G(A) - B\|_1, \mathcal{L}_{\text{L1}}^{\text{masked}} = \|M \odot (G(A) - B)\|_1, \tag{6}$$

where M is the combined mask for the detection zone, RGB anchors, and white reference, and $\odot$ denotes element-wise multiplication.

Angular Illumination Consistency Loss Hordley & Finlayson (2004) To preserve illumination consistency, we compare the angular deviation between the ground-truth illumination field and the predicted illumination field. The illumination fields are estimated as,

$$I_{\text{gt}} = \frac{A}{B + \epsilon}, \qquad I_{\text{pred}} = \frac{A}{G(A) + \epsilon}, \tag{7}$$

and the angular consistency loss is defined as,

$$\mathcal{L}_{\text{ang}} = \frac{1}{N} \sum_{p} \angle\big(I_{\text{gt}}(p), I_{\text{pred}}(p)\big). \tag{8}$$

Here, $\angle(\cdot, \cdot)$ denotes the angular error between two illumination vectors, and $\epsilon = 10^{-4}$ is used for numerical stability.

Structured Concentration Loss (NeurSPR) NeurSPR predicts continuous creatinine concentration using a bin-aware structured formulation. Training the model utilizes a bin classification loss $\mathcal{L}_{\text{bin}}$ for correct bin assignment and a regression loss $\mathcal{L}_{\text{reg}}$ Hastie et al. (2009) for minimizing the creatinine concentration prediction error within each bin,

$$\mathcal{L}_{\text{bin}} = -\sum_{k=1}^{K} t_k \log p_k, \quad \mathcal{L}_{\text{reg}} = \sum_{k=1}^{K} p_k (\hat{y}_k - y)^2, \tag{9}$$

where t_k represents soft bin targets and y is the ground-truth concentration. NeurSPR is trained on white-light scenario images and its weights are frozen during the TACC-GAN training, forcing TACC-GAN to shift the color domain of the input image to align with its white-light scenario equivalent.

Status Classification Loss The status classification loss $\mathcal{L}_{status}$ is applied as a multi-class cross-entropy objective Goodfellow et al. (2016). Given the ground-truth status probability s_c (derived from laboratory-calibrated eGFR) and the predicted status probability $\tilde{s}_c$, the loss is computed as

$$\mathcal{L}_{status} = -\sum_{c \in \mathcal{C}} s_c \log(\tilde{s}_c). \tag{10}$$

4 Evaluation Design

4.1 Dataset

Data collection utilized the Kidney Health Monitoring (KHM) mobile application framework described by Whelan et al. Whelan et al. (2024). The core dataset originates from 504 physical creatinine test strips, where biochemical concentrations were laboratory-prepared to guarantee a rigorous, controlled ground-truth range for evaluation.

US Population-Based Sampling: We assign demographic metadata (age, sex, race) from 2019 US Census statistics, as the eGFR-derived label depends non-linearly on these factors.

Strict Patient-Wise Split: To rigorously evaluate generalization and prevent data leakage, each of the 504 unique strips is assigned to a distinct patient ID, and the training, validation, and test sets are completely disjoint at the patient level. Thus no image (under any lighting) from a test-set patient appears in training, forcing the model to learn true color constancy rather than memorizing strip-specific textures. No augmentation is applied to the validation or test sets; the 404 test images are full, un-augmented strips from 101 held-out patients.

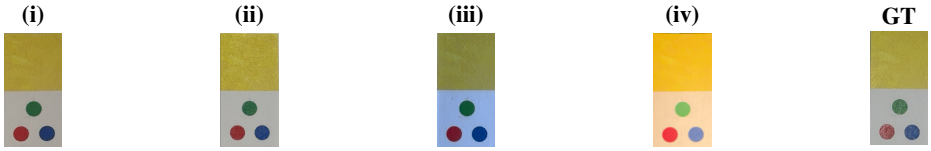


Figure 2: Representative dataset samples across lighting scenarios (i) Indoor warm light, (ii) Indoor white light, (iii) Outdoor natural light, (iv) Random light variations, and Ground-truth white-light reference.

Environmental Diversity: Each patient sample was captured under four distinct lighting environments: (1) Indoor-warm, (2) Indoor-white, (3) Outdoor-natural, and (4) Random variations, to simulate unconstrained at-home testing. This yields exactly 2,016 images (504×4) for the color-correction stage, split at the patient level into 1,528 training, 84 validation, and 404 test images ($382/21/101$ patients $\times 4$). The diagnostic module (NeurSPR) is trained, validated and tested separately on the 504 white-light strips under the same patient split and they also serve as ground truth targets for color correction before final prediction.

4.2 Baselines and Comparison Setup

Color Constancy Baselines. We compare TACC-GAN against three families of correction methods. *(i) Learned image-to-image*: Deep White-Balance (DeepWB) (Afifi & Brown, 2020) and AngularGAN (Sidorov, 2019), our primary generative baseline. *(ii) Reference-free statistical*: Gray-World (Buchsbaum, 1980), White-Patch (Land, 1977), and Shades-of-Gray (Finlayson & Trezzi, 2004). *(iii) Reference-based oracle*: Von Kries diagonal correction (von Kries, 1970) and a 3×3 Color Correction Matrix, CCM (Finlayson et al., 2015), which use the strip's white patch and paired white-light reference at inference; these are not deployable in unconstrained settings and serve as an upper bound on achievable calibration. We omit global illuminant estimators (FC4 (Hu et al., 2017), C4 (Yu et al., 2020)), as a single 1×3 vector cannot correct the spatially varying ISP non-linearities present in mobile captures.

Clinical Baselines (Diagnostic Predictors). To isolate the impact of NeurSPR, we compare against the SPR (Whelan et al., 2024) pipeline with non-differentiable classifiers (Logistic Regression, Decision Trees, Random Forest, Histogram Boosted Trees) and a direct CNN-regression baseline that replaces the KDE histogram layer with spatial CNN features under the same diagnostic losses.

4.3 Evaluation Metrics

Our evaluation prioritizes clinical utility over perceptual image quality. For **kidney health classification**, we report weighted Precision, Recall, and weighted F1-score (our primary metric for class imbalance across Healthy, Intermediate, and Critical classes), alongside creatinine concentration RMSE (mg/dL) at the NeurSPR output. For **biochemical color fidelity**, we compute KDE overlap in Lab color space—specifically the b-channel capturing the Jaffe reaction's yellow-blue shift against the white-light reference—rather than background-dominated metrics like PSNR or SSIM.

4.4 Implementation Details

All models are implemented in PyTorch and trained on NVIDIA GPUs, including A6000 24GB, V100 32GB, A100 80GB, and GH200 96GB. TACC-GAN and NeurSPR are trained end-to-end using Adam ($\beta_1 = 0.5, \beta_2 = 0.999$) with a learning rate of 2×10^{-4}. Both TACC-GAN variants use $\lambda_{\text{GAN}} = 0.25$, $\lambda_{\text{L1}} = 0.10$, $\lambda_{\text{L1,global}} = 0.15$, and $\lambda_{\text{L1,masked}} = 0.25$. TACC-GAN$_1$ uses $\lambda_{\text{ang}} = 0.20$ and $\lambda_{\text{NeurSPR}} = 0.22$, while TACC-GAN$_2$ uses 0.15 and 0.20, respectively. The NeurSPR regression experts use the SPR Whelan et al. (2024) concentration intervals $[0, 2.8)$, $[2.8, 5.7)$, $[5.7, 19.7)$, and $[19.7, 60.0]$. Code and data will be released publicly upon acceptance.

5 Experimental Results

5.1 Kidney Health Status Classification Performance

Table 1 reports classification performance and concentration RMSE across the four lighting conditions and an aggregated "All Scenarios" setting (unknown illumination). Without correction, per-

Table 1: **Kidney Health Status Classification Performance Across Lighting Scenarios.** Methods are grouped into *deployable* (reference-free at inference) and *oracle* (requires the paired white-light reference at test time and reported as an upper bound, not a competitor).

Method	Indoor Warm				Indoor White				Outdoor Natural				Random Variations				All Scenarios			
	P	R	F1	RMSE	P	R	F1	RMSE	P	R	F1	RMSE	P	R	F1	RMSE	P	R	F1	RMSE
Deployable (reference-free at inference)																				
No CC	0.6614	0.6139	0.5809	10.4128	0.8392	0.8317	0.8284	5.5528	0.7383	0.6436	0.5905	15.9802	0.5099	0.5248	0.5050	14.2827	0.6524	0.6535	0.6343	12.2334
Deep-WB	0.8043	0.7723	0.7764	**4.9131**	0.7682	0.7327	0.7380	5.4666	0.7522	0.7129	0.7186	5.8448	0.7317	0.6634	0.6761	6.9580	0.7637	0.7203	0.7273	5.8437
AngularGAN	0.8455	0.8317	0.8334	5.8654	0.8256	0.8020	0.8034	**4.9180**	0.8381	0.8317	0.8319	**4.9036**	0.8330	0.8119	0.8144	**5.0393**	0.8341	0.8193	0.8208	**5.1969**
Shades of Gray	0.6397	0.4158	0.3218	13.7219	0.6427	0.4356	0.3597	13.3058	0.3612	0.5050	0.3794	14.5546	0.1487	0.2772	0.1809	14.2541	0.6370	0.4084	0.3187	13.9674
White Patch	0.6389	0.4158	0.3195	13.8547	0.6578	0.4455	0.3454	13.3045	0.3485	0.5248	0.4017	14.2405	0.6364	0.4059	0.3135	14.2824	0.6624	0.4480	0.3462	13.9261
Gray World	0.5810	0.3564	0.3047	13.2586	0.5757	0.3465	0.2874	13.5670	0.7520	0.5644	0.5086	12.0323	0.1470	0.2772	0.1813	14.2349	0.6103	0.3861	0.3354	13.2972
TACC-GAN$_1$	**0.8908**	**0.8812**	**0.8818**	8.6604	**0.8779**	**0.8713**	**0.8719**	7.1515	**0.9147**	**0.9010**	**0.9016**	8.3260	**0.8756**	0.8614	0.8625	8.2553	**0.8884**	**0.8787**	**0.8795**	8.1181
TACC-GAN$_2$	0.8734	0.8713	0.8709	5.4030	0.8742	0.8713	0.8718	5.7565	0.8936	0.8911	0.8916	5.5524	0.8728	**0.8713**	**0.8717**	5.1524	0.8777	0.8762	0.8766	5.4705
***Oracle (paired white-light reference required at inference; reported as upper bound and not deployable)*†**																				
Von Kries†	0.8812	0.8812	0.8808	4.6837	0.8728	0.8713	0.8716	4.7465	0.8868	0.8812	0.8813	6.4370	0.8026	0.8020	0.8017	5.7247	0.8600	0.8589	0.8591	5.4468
CCM†	0.8263	0.8119	0.8124	7.5965	0.7575	0.7426	0.7372	6.9762	0.8434	0.8317	0.8323	6.6304	0.7646	0.7525	0.7478	6.9982	0.7979	0.7847	0.7828	7.0589

Note: P, R, F1, and RMSE denote Precision, Recall, F1-score, and root-mean-square error of predicted creatinine concentration in mg/dL (lower is better), respectively. All metrics are computed at the NeurSPR output after each correction method. †Von Kries and CCM require the paired white-light reference image at inference time and are therefore not deployable in unconstrained settings.

formance collapses under the hardest *Random-Variations* scenario (F1 = 0.5050), confirming that unconstrained lighting destroys the colorimetric signal of the Jaffe reaction. Deep-WB and AngularGAN both improve over this baseline, though Deep-WB (All-Scenarios F1 = 0.7273) struggles with spatially varying ISP non-linearities. The global-statistical methods (Gray-World, Shades-of-Gray) perform poorly, achieving All-Scenarios F1 scores of 0.3354 and 0.3187, respectively, as their gray/white-balance assumptions are violated by the strip's saturated colored background.

We report two variants that differ only in loss weighting (Section 4.4): the headline **TACC-GAN$_2$**, which balances classification and fidelity, and **TACC-GAN$_1$**, a variant that maximizes F1. Among deployable methods, TACC-GAN achieves the best classification in every scenario: TACC-GAN$_2$ reaches an All-Scenarios F1 of 0.8766 (+38.2% over No CC, +0.0558 over AngularGAN) at RMSE 5.47 mg/dL, and lifts Random-Variations F1 from 0.5050 to 0.8717 (+72.6% relative). Crucially, it is comparable to the non-deployable oracle Von Kries on both F1 (0.8766 vs. 0.8591) and RMSE (5.47 vs. 5.45) despite using no reference at test time, recovering most of the benefit of explicit calibration in a deployable form. The divergence between TACC-GAN$_1$'s high F1 and higher RMSE suggests that its gains are concentrated around clinically relevant decision boundaries, whereas TACC-GAN$_2$ provides a better balance between status classification and continuous concentration recovery.

5.2 Qualitative and Chemical Fidelity Analysis

Beyond perceptual quality, diagnostic accuracy requires preserving the specific spectral signature of the underlying biochemistry. As shown in Figure 3, baselines like DeepWB and AngularGAN struggle with spatially varying illumination, leading to subtle hue distortions and flattened chromatic peaks in a/b channels under random lighting (Figures 4a and 4b). In contrast, **TACC-GAN** closely matches the white-light ground truth across all lighting scenarios. Kernel Density Estimation (KDE) in *Lab* color space confirms that TACC-GAN recovers sharp peaks in the b-channel (capturing the Jaffe reaction's yellow-blue shift) aligned with ground truth, proving that NeurSPR's task-aware feedback prioritizes chemically valid radiometric profiles over generic white balance.

5.3 NeurSPR vs. Classical SPR

We validate NeurSPR against the original SPR pipeline with non-differentiable classifiers (Logistic Regression, Decision Trees, Support Vector Machines, Histogram Boosted Trees). As shown in Table 2, NeurSPR achieves the best F1 (0.9212) versus the strongest classical baseline (Histogram Boosted Trees, 0.9112). While modest, this is not the point: NeurSPR's differentiability lets diagnostic losses propagate into TACC-GAN during task-aware correction which is impossible with classical SPR.

6 Ablation Study: Dissecting the Gains

We conduct two ablations: the aggregation mechanism within NeurSPR, and the contributions of spatial masking and task-aware feedback within TACC-GAN. First, we compare two strategies for

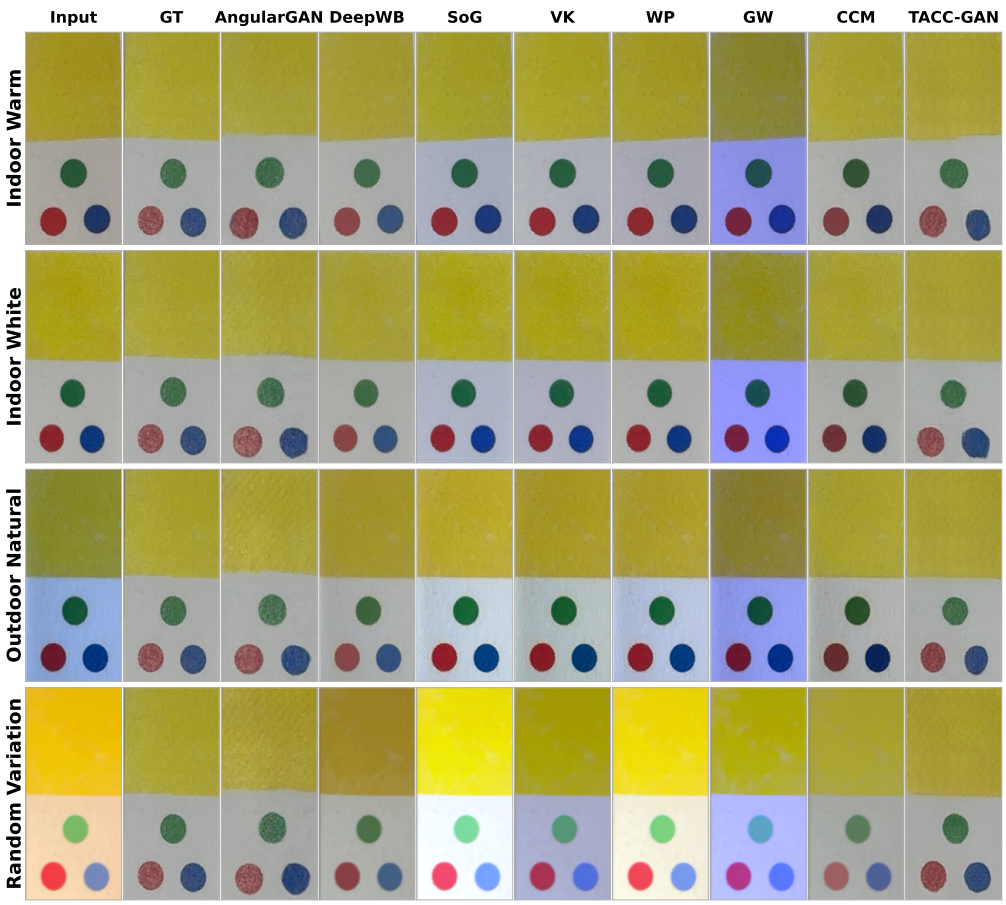


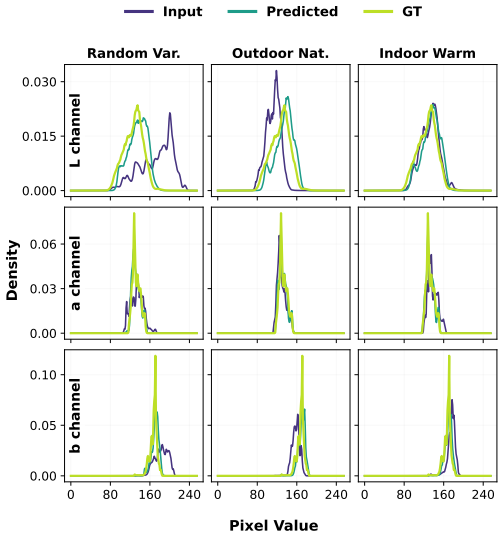
Figure 3: **Qualitative comparison of color constancy methods across different lighting scenarios for a single test strip sample.**

Table 2: CKD Classification Performance of SPR and NeurSPR Models

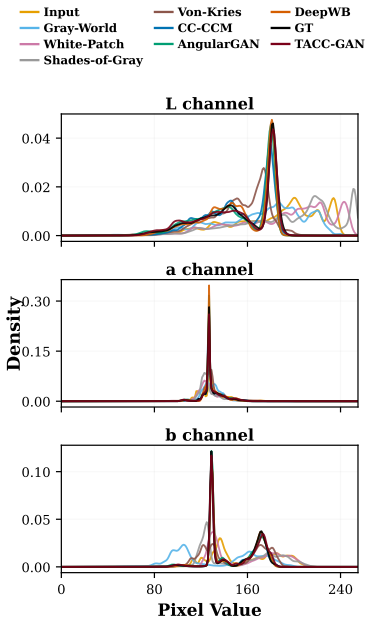
Method	Precision	Recall	F1-score	RMSE
SPR (Logistic Regression)	0.8597	0.8416	0.8441	6.7547
SPR (Decision Trees)	0.8732	0.8713	0.8700	5.7269
SPR (Support Vector Machine)	0.8565	0.0.8515	0.8521	**4.3762**
SPR (Histogram Boosted Trees)	0.9133	0.9109	0.9112	5.0566
SPR (Direct CNN)	0.8700	0.8190	0.8437	6.3511
NeurSPR (Ours)	**0.9225**	**0.9208**	**0.9212**	6.2661

aggregating NeurSPR's regression experts: *Hard Bin Selection (Argmax)*, where only the single highest-probability expert produces the concentration, and *Soft Mixture-of-Experts*, where the concentration is the softmax-probability-weighted sum of all experts. Hard selection introduces discontinuities at bin boundaries that can destabilize the gradient flow required for end-to-end training. As shown in Table 3, the **Soft Mixture** strategy outperforms Hard Bin Selection (F1 0.9212 vs. 0.9129), confirming that differentiability in the aggregation layer is critical for end-to-end optimization.

Second, we evaluate three TACC-GAN configurations to separate spatial attention from clinical alignment: *Raw* (global correction only); *Masks* (adds spatial supervision over the detection zone, RGB anchors, and white reference); and *Masks + Feedback* (further adds diagnostic feedback from NeurSPR). The results in Table 4 reveal a clear hierarchy. Adding spatial masks improves performance in controlled lighting (e.g., Indoor Warm F1 increases from 0.8244 to 0.8916), confirming that constraining the generator to the Detection Zone prevents it from wasting capacity on irrelevant background artifacts; however, masks alone do not help under harder shifts (e.g., Outdoor Natural F1 drops from 0.8913 to 0.8153, and Random Variations remains at 0.7646). Task-aware feedback is the



(a) KDE of LAB distributions for **TACC-GAN**. Columns are lighting conditions (L-R: Random, Outdoor, Indoor-Warm).



(b) Comparison of LAB distributions under the 'Random-Variations' scenario.

Figure 4: LAB color distribution analysis. (a) TACC-GAN across lighting conditions; (b) method comparison across baselines.

Table 3: **Ablation on NeurSPR Aggregation Strategy**

Method	**Precision**	**Recall**	**F1-score**	**RMSE**
NeurSPR (Hard Argmax)	0.9226	0.9074	0.9129	**5.9877**
NeurSPR (Soft Mixture)	**0.9225**	**0.9208**	**0.9212**	6.2661

Note: NeurSPR predicts CKD classes. All variants share the same backbone and bin classifier.

Table 4: **Ablation of TACC-GAN Components Across Lighting Scenarios.**

Method	**Indoor Warm**				**Indoor White**				**Outdoor Natural**				**Random Variations**				**All Scenarios**			
	P	**R**	**F1**	**RMSE**	**P**	**R**	**F1**	**RMSE**	**P**	**R**	**F1**	**RMSE**	**P**	**R**	**F1**	**RMSE**	**P**	**R**	**F1**	**RMSE**
Raw	0.8410	0.8218	0.8244	**5.4728**	0.8264	0.7921	0.7956	6.7069	0.8945	0.8911	0.8913	6.0055	0.7934	0.7426	0.7471	8.5066	0.8346	0.8119	0.8144	6.7706
Masks	**0.8941**	**0.8911**	**0.8916**	5.7926	0.8548	0.8416	0.8434	**3.4072**	0.8390	0.8119	0.8153	**5.2547**	0.7885	0.7624	0.7646	**5.1207**	0.8416	0.8267	0.8288	**4.9749**
Masks + Feedback	0.8908	0.8812	0.8818	8.6604	**0.8779**	**0.8713**	**0.8719**	7.1515	**0.9147**	**0.9010**	**0.9016**	8.3260	**0.8756**	**0.8614**	**0.8625**	8.2553	**0.8884**	**0.8787**	**0.8795**	8.1181

Note: P, R, F1, and RMSE denote Precision, Recall, F1-score, and root-mean-square error of predicted creatinine concentration respectively.

decisive factor for robustness in unconstrained environments: in the challenging *Random Variations* scenario, adding feedback raises F1 from 0.7646 (Masks only) to 0.8625 (Full Model), a **+9.8% absolute** gain. While masking ensures spatial fidelity, the diagnostic feedback loop is essential for learning the precise transformations required for clinical accuracy under complex lighting.

7 Conclusion

We presented TACC, a task-aware color constancy framework for smartphone-based CKD prediction under uncontrolled lighting. By coupling TACC-GAN with differentiable NeurSPR, diagnostic losses can directly guide color correction, overcoming the non-differentiable bottleneck of classical SPR pipelines. Experiments across indoor, outdoor, and random lighting show that task-aware feedback improves clinical robustness while preserving the chromatic cues required for creatinine estimation. Future work will expand the dataset with additional patient samples to improve coverage of clinical variability and further evaluate the generalizability of the proposed framework. Beyond CKD monitoring, the same paradigm can extend to other colorimetric point-of-care diagnostics relying on precise colorimetric analysis, such as urinalysis, blood glucose monitoring, or pH detection.

REFERENCES

Mahmoud Afifi and Michael S. Brown. Deep white-balance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Gershon Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin Institute*, 310(1):1–26, 1980. doi: 10.1016/0016-0032(80)90058-7.

Centers for Disease Control and Prevention. Chronic Kidney Disease in the United States, 2023. Technical report, US Department of Health and Human Services, 2023. URL `https://www.cdc.gov/kidney-disease/php/data-research/index.html`.

Graham D. Finlayson and Elisabetta Trezzi. Shades of gray and colour constancy. In *Proceedings of the 12th Color Imaging Conference*, pp. 37–41, 2004. doi: 10.2352/CIC.2004.12.1.ART00008.

Graham D. Finlayson, Michal Mackiewicz, and Anya Hurlbert. Colour correction using root-polynomial regression. *IEEE Transactions on Image Processing*, 24(5):1460–1470, 2015. doi: 10.1109/TIP.2015.2405336.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.

Steven D Hordley and Graham D Finlayson. A re-evaluation of color constancy algorithm performance. *Journal of the Optical Society of America A*, 21(2):285–297, 2004.

Yuanming Hu, Yangyan Chen, Stephen Luo, and Ariel Shamir. FC4: Fully Convolutional Color Constancy with Confidence-Weighted Pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Hakki Can Karaimer and Michael S Brown. A software platform for manipulating the camera imaging pipeline. In *European Conference on Computer Vision*, pp. 429–444. Springer, 2016.

LabX. Shop Clinical Chemistry Equipment, New and Used Prices. `https://www.labx.com/categories/clinical-chemistry`, 2025. Accessed: Feb 2026.

Edwin H. Land. The retinex theory of color vision. *Scientific American*, 237(6):108–128, 1977. doi: 10.1038/scientificamerican1277-108.

Andrew S Levey, Josef Coresh, Tom Greene, Jane Marsh, Lesley A Stevens, John W Kusek, Frederick Van Lente, and Chronic Kidney Disease Epidemiology Collaboration. Expressing the modification of diet in renal disease study equation for estimating glomerular filtration rate with standardized serum creatinine values. *Clinical chemistry*, 53(4):766–772, 2007.

Zhi-Pei Liang and Paul C Lauterbur. Constrained reconstruction methods in mr imaging. *IEEE Reviews in Biomedical Engineering*, 2000.

Joshua Peeples, Weihuang Xu, and Alina Zare. Histogram layers for texture analysis. *IEEE Transactions on Artificial Intelligence*, 3(4):541–552, 2021.

Amirreza Shaban, Christoph Baugh, Dong Lao Kang, and Caie Maxwell. Staingan: Stain style transfer for digital histological images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 953–956. IEEE, 2019.

Oleksii Sidorov. Conditional gans for multi-illuminant color constancy: Revolution or yet another approach? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

Johannes von Kries. Influence of adaptation on the effects produced by luminous stimuli. In David L. MacAdam (ed.), *Sources of Color Science*. MIT Press, Cambridge, MA, 1970.

Zhe Wang, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learnable histogram: Statistical context features for deep neural networks. In *European Conference on Computer Vision (ECCV)*, pp. 246–262. Springer, 2016.

Alex Whelan, Soham Phadke, Alessandro Bellofiore, and David C. Anastasiu. On-device prediction for chronic kidney disease. In *2022 IEEE Global Humanitarian Technology Conference (GHTC)*, pp. 325–332, 2022. doi: 10.1109/GHTC55712.2022.9910606.

Alex Whelan, Ragwa Elsayed, Alessandro Bellofiore, and David C Anastasiu. Selective partitioned regression for accurate kidney health monitoring. *Annals of Biomedical Engineering*, 52(5):1448–1462, 2024.

Huanglin Yu, Ke Chen, Kaiqi Wang, Yanlin Qian, Zhaoxiang Zhang, and Kui Jia. Cascading convolutional color constancy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12725–12732, 2020.

Xingguang Zhang et al. Vision-Driven Image Signal Processing: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024. Standard reference for task-aware image enhancement paradigms.

Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016.