# A Appendix

## A.1 Data statistics

Table A1: Detailed statistics of data splits across 10 random seeds on the CNUH-CKD dataset, including the number of examinations, images, and samples used in each subset.

| Split | Training set | | | Validation set | | | Test set | | |
|---|---|---|---|---|---|---|---|---|---|
| | Examinations | Images | Samples | Examinations | Images | Samples | Examinations | Images | Samples |
| 1 | 906 | 7909 | 7909 | 300 | 2667 | 2667 | 302 | 2669 | 1127 |
| 2 | 908 | 7957 | 7957 | 300 | 2659 | 2659 | 300 | 2629 | 1113 |
| 3 | 903 | 8004 | 8004 | 302 | 2588 | 2588 | 303 | 2653 | 1111 |
| 4 | 904 | 8031 | 8031 | 303 | 2648 | 2648 | 301 | 2566 | 1077 |
| 5 | 906 | 8010 | 8010 | 302 | 2577 | 2577 | 300 | 2658 | 1113 |
| 6 | 904 | 7865 | 7865 | 303 | 2794 | 2794 | 301 | 2586 | 1082 |
| 7 | 908 | 8042 | 8042 | 299 | 2641 | 2641 | 301 | 2562 | 1051 |
| 8 | 904 | 7978 | 7978 | 299 | 2647 | 2647 | 305 | 2620 | 1080 |
| 9 | 904 | 8046 | 8046 | 306 | 2639 | 2639 | 298 | 2560 | 1074 |
| 10 | 905 | 8047 | 8047 | 300 | 2629 | 2629 | 303 | 2569 | 1077 |

## A.2 Traditional machine learning models

To investigate the effectiveness of USFM-derived feature representations, we trained three traditional machine learning classifiers (Logistic Regression (LR) [31], Support Vector Machine (SVM) [32], XGBoost [33]) on the extracted embeddings from the USFM encoder, with class weights incorporated during training. The results are shown in Table A2.

Table A2: Performance metrics (%) of machine learning models using foundation model embeddings at both the sample and examination levels across 10 Monte Carlo cross-validation runs.

| | **Model** | **B-ACC** | **F1** | **AUC** | **Sensitivity** | **Specificity** |
|---|---|---|---|---|---|---|
| Sample level | USFM+LR | $71.88 \pm 1.80$ | $68.95 \pm 1.54$ | $79.23 \pm 2.17$ | $76.49 \pm 2.29$ | $67.26 \pm 4.51$ |
| | USFM+SVM | $65.86 \pm 2.52$ | $67.44 \pm 2.72$ | $81.41 \pm 2.55$ | $91.19 \pm 1.26$ | $40.53 \pm 4.84$ |
| | USFM+XGBoost | $67.39 \pm 2.81$ | $67.87 \pm 2.79$ | $78.69 \pm 2.52$ | $86.20 \pm 1.21$ | $48.59 \pm 5.45$ |
| Examination level | USFM+LR | $75.50 \pm 2.92$ | $72.02 \pm 2.03$ | $82.31 \pm 1.85$ | $78.81 \pm 2.43$ | $72.17 \pm 7.02$ |
| | USFM+SVM | $66.54 \pm 2.95$ | $68.38 \pm 3.20$ | $83.11 \pm 2.12$ | $92.26 \pm 0.92$ | $40.83 \pm 5.41$ |
| | USFM+XGBoost | $68.18 \pm 3.47$ | $69.16 \pm 3.59$ | $81.27 \pm 2.04$ | $88.82 \pm 1.19$ | $47.54 \pm 6.26$ |

Abbreviations: B-ACC, balanced accuracy; F1, F1-score; AUC, area under the receiver operating characteristic curve.

Among the three classifiers trained on USFM-derived embeddings, LR achieved the most balanced performance, with the highest B-ACC ($71.88 \pm 1.80\%$ at the sample level and $75.50 \pm 2.92\%$ at the examination level) and the highest specificity ($67.26 \pm 4.51\%$ and $72.17 \pm 7.02\%$, respectively), at the cost of lower sensitivity. In contrast, USFM+SVM and USFM+XGBoost achieved higher sensitivity (above 86% at the sample level and above 88% at the examination level) but substantially lower specificity (around 40–48%), reflecting the same tendency to over-predict the positive class observed for EfficientNet-B0 in the main analysis.

## A.3 Patient-level performance

For the patient-level evaluation, predictions were further aggregated across all examinations belonging to the same patient. Let $P_p = \{e \mid \text{patient}(e) = p\}$ denote the set of examination indices belonging to patient $p$. The patient-level probability vector was obtained by averaging the examination-level probability vectors

within $P_p$:

$$\hat{p}_p^{\text{patient}} = \frac{1}{|P_p|} \sum_{e \in P_p} \hat{p}_e^{\text{exam}}. \tag{5}$$

The final patient-level predicted class was then obtained as:

$$\hat{y}_p^{\text{patient}} = \underset{c \in \{1,\ldots,C\}}{\arg\max}\, \hat{p}_{p,c}^{\text{patient}}. \tag{6}$$

Because the great majority of patients in the CNUH-CKD cohort underwent a single examination (1508 examinations across 1484 patients), the patient-level results are closed to the examination-level results.

As shown in Table A3, USFM achieved the strongest overall performance at the patient level, obtaining the highest B-ACC ($74.54 \pm 4.56\%$), F1-score ($73.69 \pm 3.57\%$), AUC ($84.88 \pm 1.91\%$) and specificity ($63.24 \pm 10.67\%$). VGG16 ranked second in B-ACC ($73.32 \pm 4.40\%$), AUC ($84.32 \pm 1.52\%$) and specificity ($60.71 \pm 12.95\%$), while ViT16 ranked second in F1-score ($72.55 \pm 3.14\%$) and sensitivity ($89.20 \pm 3.57\%$). EfficientNet-B0 again recorded the highest sensitivity ($91.37 \pm 2.22\%$) alongside the lowest specificity ($46.06 \pm 6.09\%$), reflecting the same tendency to over-predict the positive class observed at the sample and examination levels.

Table A3: Comparison of performance metrics (%) at the patient level across 10 Monte Carlo cross-validation runs. **Bold**: best results, underline: second-best results.

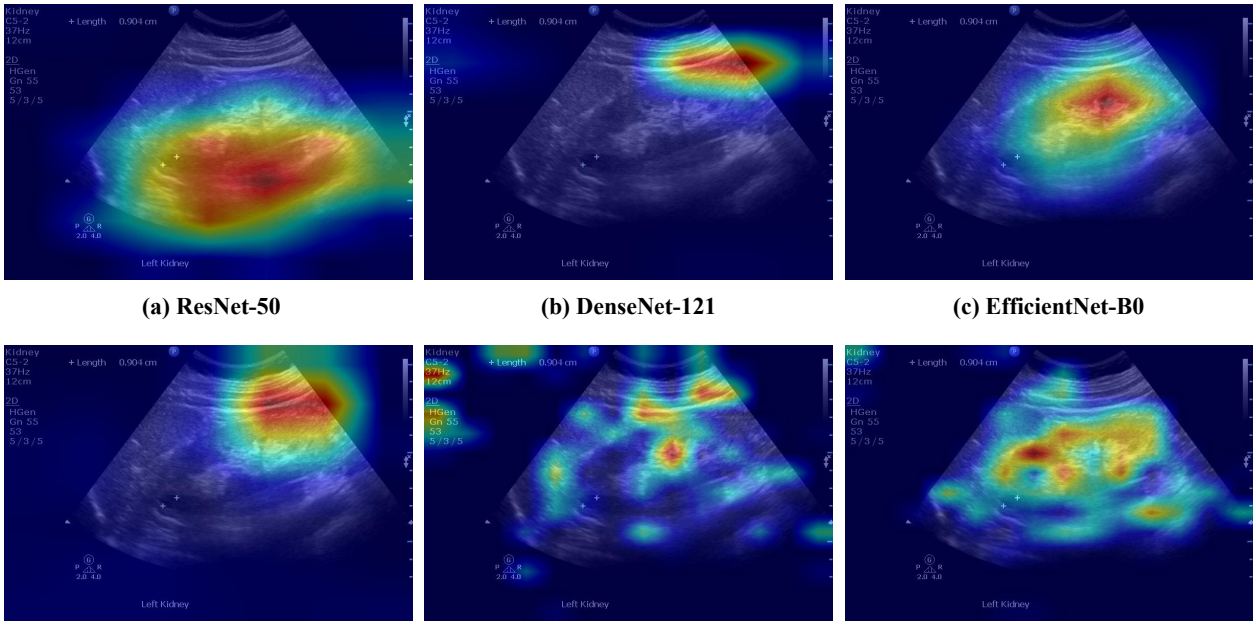
| Model | Params | B-ACC | F1 | AUC | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| ResNet-50 [21] | 23.5M | $70.71 \pm 3.50$ | $71.03 \pm 2.70$ | $82.46 \pm 2.72$ | $87.61 \pm 2.83$ | $53.81 \pm 8.90$ |
| DenseNet-121 [22] | 6.9M | $70.41 \pm 4.03$ | $70.80 \pm 2.88$ | $83.08 \pm 0.68$ | $88.01 \pm 4.24$ | $52.82 \pm 11.31$ |
| EfficientNet-B0 [23] | 4M | $68.72 \pm 2.74$ | $70.34 \pm 2.68$ | $82.72 \pm 1.84$ | $\mathbf{91.37 \pm 2.22}$ | $46.06 \pm 6.09$ |
| VGG16 [24] | 134.3M | $\underline{73.32 \pm 4.40}$ | $72.52 \pm 3.17$ | $\underline{84.32 \pm 1.52}$ | $85.93 \pm 4.81$ | $\underline{60.71 \pm 12.95}$ |
| ViT16 [25] | 85.8M | $71.93 \pm 3.87$ | $\underline{72.52 \pm 3.14}$ | $84.20 \pm 2.15$ | $\underline{89.20 \pm 3.57}$ | $54.65 \pm 10.41$ |
| USFM [18] | **39K** | $\mathbf{74.54 \pm 4.56}$ | $\mathbf{73.69 \pm 3.57}$ | $\mathbf{84.88 \pm 1.91}$ | $85.84 \pm 3.45$ | $\mathbf{63.24 \pm 10.67}$ |

Abbreviations: B-ACC, balanced accuracy; F1, F1-score; AUC, area under the receiver operating characteristic curve.

Overall, the patterns observed at the patient level are consistent with those reported at the sample and examination levels, with USFM maintaining the most favourable balance between sensitivity and specificity across all evaluation granularities. This consistency indicates that the diagnostic advantage of the ultrasound foundation model is preserved regardless of the aggregation strategy and that the proposed approach yields stable performance when predictions are summarised at the level of the individual patient.

### A.4 Visualization

Figure A.1 presents XGrad-CAM [34] visualizations for six models on a representative CKD sample. These visualizations are provided as qualitative examples to illustrate potential differences in model attention, rather than as evidence of consistent behavior across the full dataset. In the examples shown, CNN-based models produced relatively broad activation regions, which may indicate reliance on general renal image patterns rather than specific anatomical subregions. DenseNet-121 showed activation concentrated in the upper right region of the image, outside the renal parenchyma, whereas VGG16 exhibited a partially shifted activation toward the upper area, with limited overlap with the kidney. ViT16 showed dispersed activations extending beyond the kidney boundary, suggesting that the highlighted regions in this case were not well localized to clinically relevant renal structures despite its competitive classification performance. By comparison, USFM showed a more distributed activation pattern, with relatively stronger responses around the renal cortex. Given that the renal cortex is associated with CKD-related structural alterations, such as cortical thinning and fibrosis [35], this example suggests that ultrasound-specific pretraining may encourage attention to more anatomically meaningful regions. Nevertheless, these findings should be interpreted cautiously, as visual explanations from selected cases cannot establish dataset-wide model behavior.



**(a) ResNet-50** **(b) DenseNet-121** **(c) EfficientNet-B0**

**(d) VGG16** **(e) ViT16** **(f) USFM**

Figure A.1: Class activation map generated using XGradCAM for (a) `ResNet-50`, (b) `DenseNet-121`, (c) `EfficientNet-B0`, (d) `VGG16`, (e) `ViT16` and (f) `USFM` on a representative sample, visualizing the important regions that influenced the model's prediction for the CKD class.