# Solving the Reference Pathologist Paradox in Machine Learning Development for Histology Scoring

Thomas Forest[1], Sabu Kuruvilla[1], Binod Jacob[1], Nagaraja Muniappa[1], Takayuki Tsuchiya[1], Raymond Gonzalez[1], Malini Roy[2], Raghav Amaravadi[2], Geetank Raipuria[2], Nitin Singhal[2]

[1]Merck & Co., Inc., Rahway NJ, USA
[2]Aira Matrix, Mumbai, India

## Abstract

**Introduction** – Development of machine learning (ML) algorithms for scoring histology slides commonly involves training against example histopathology findings (Supervised Learning). This approach creates a ML performance boundary based on the list of diagnoses included and the observations recorded by the reference pathologists. We hypothesized that a ML development strategy not requiring training against histopathology findings (Unsupervised Learning) could increase algorithm performance by identifying novel findings.

**Design** – Two ML algorithms were developed for scoring Han Wistar rat kidney histology. ML scoring was compared to the independent evaluations of 3 experienced toxicologic pathologists using a rat study of carbapenem, a classical renal toxicant.

**Results** – Scores from the ML trained using examples of renal tubular histopathology aligned closely with the consensus of the pathologist panel. Whereas scores from a ML trained only using histology from vehicle treated rats identified a subtle histomorphology difference in a dose group anticipated to be not remarkable based on previous studies and considered not remarkable by the consensus of the pathologist panel.

**Conclusion** – A ML algorithm that scored histology based on deviation from a model of normal histomorphology identified a subtle non-adverse difference between control and treated groups that ML trained using histopathology examples did not identify. These differences were not considered toxicologically noteworthy by a panel of experienced pathologists.

**Impact** – Advances in ML development for scoring histology slides introduce a novel frontier for detecting subtle histomorphology differences in nonclinical toxicology studies that may need to be incorporated into risk assessments in future workflows.

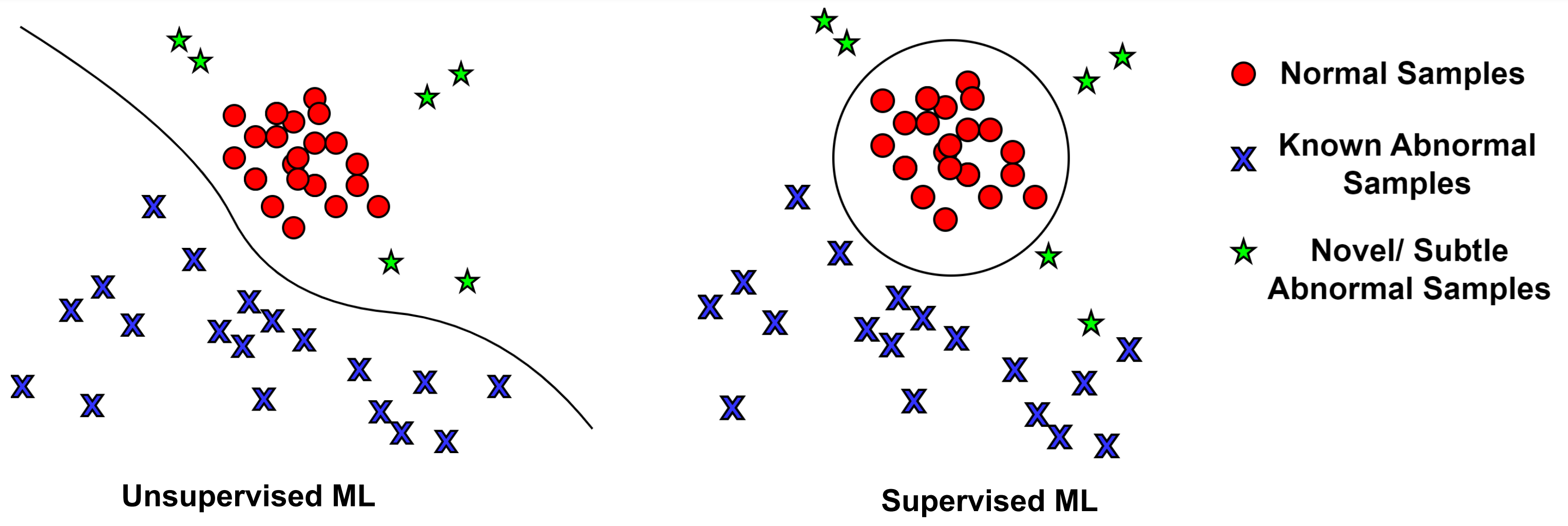

Figure 1. A comparison of supervised (right) and unsupervised (left) training strategies for machine learning. With supervised strategies the ML is unlikely to identify novel findings. Unsupervised strategies identify novel findings but may lack sensitivity.

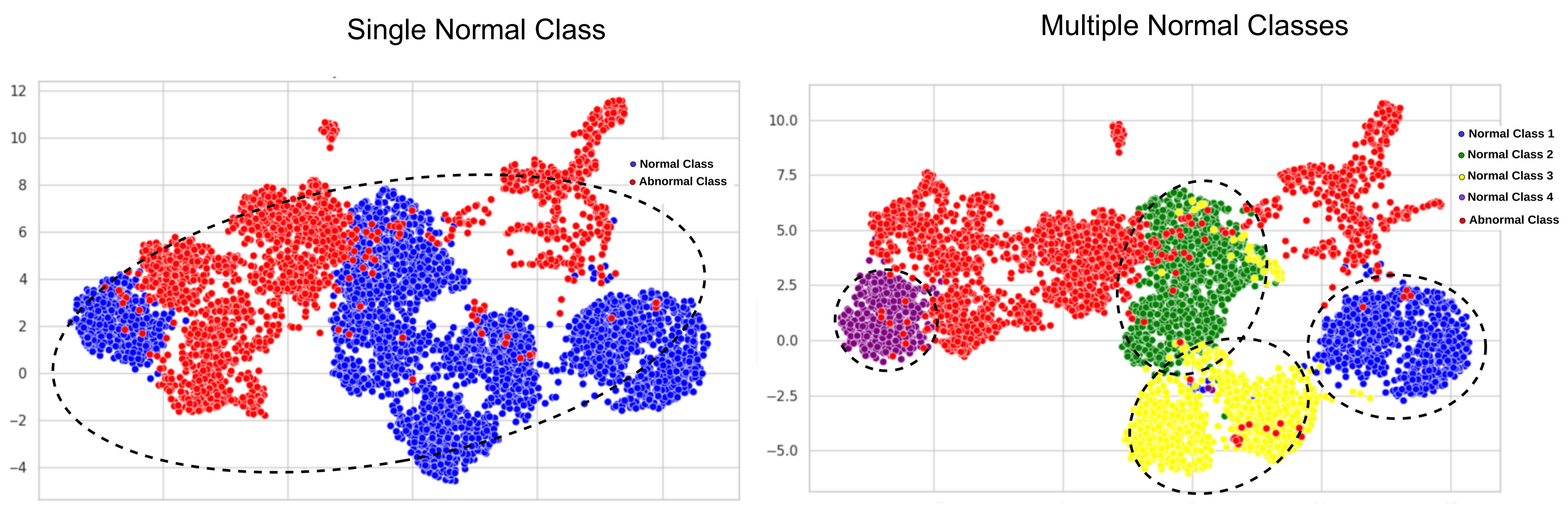

Figure 2. Some sensitivity and specificity limitations of unsupervised strategies can be addressed by subdivision of normal histology (right) where the feature space contains distinct morphologic domains.

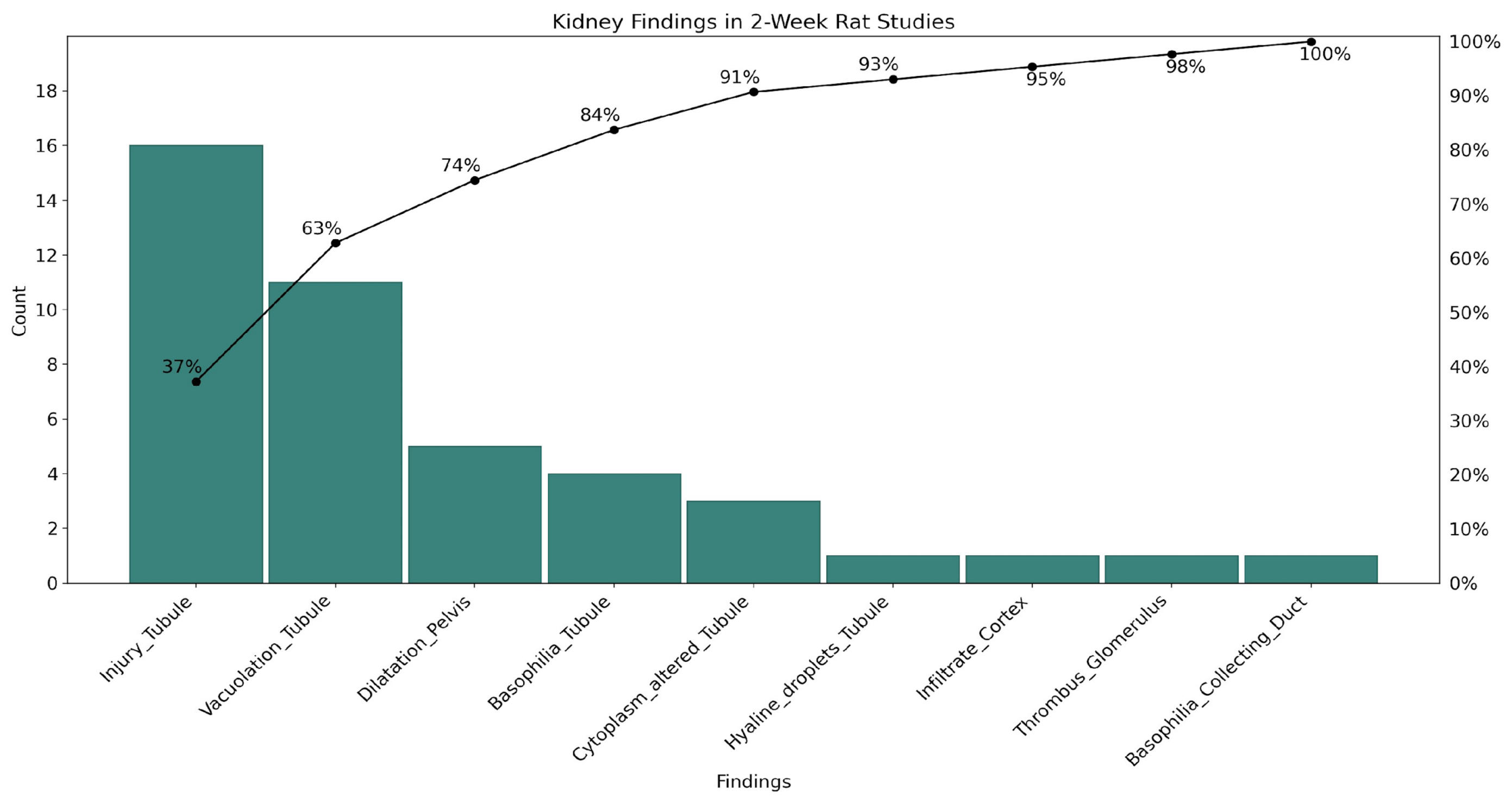

Figure 3. Incidence of histopathology findings in 2-week rat studies (y-axis = # studies) over more than 20 years. Approximately ¾ of the findings are defined by 3 diagnoses, but there are many low incidence findings as well. Such low incidence findings may appear novel in unsupervised ML strategies (Analysis and image curtesy of Mel Dsouza).

Carbapenem Study Design

| Group | Dose (mg/kg/day) | Females | Males |
|---|---|---|---|
| Control | 0 | 3 | 3 |
| Low | 75 | 5 | 5 |
| Mid | 150 | 5 | 5 |
| High | 225 | 5 | 5 |

Note: 3 pathologists independently evaluated all animals.

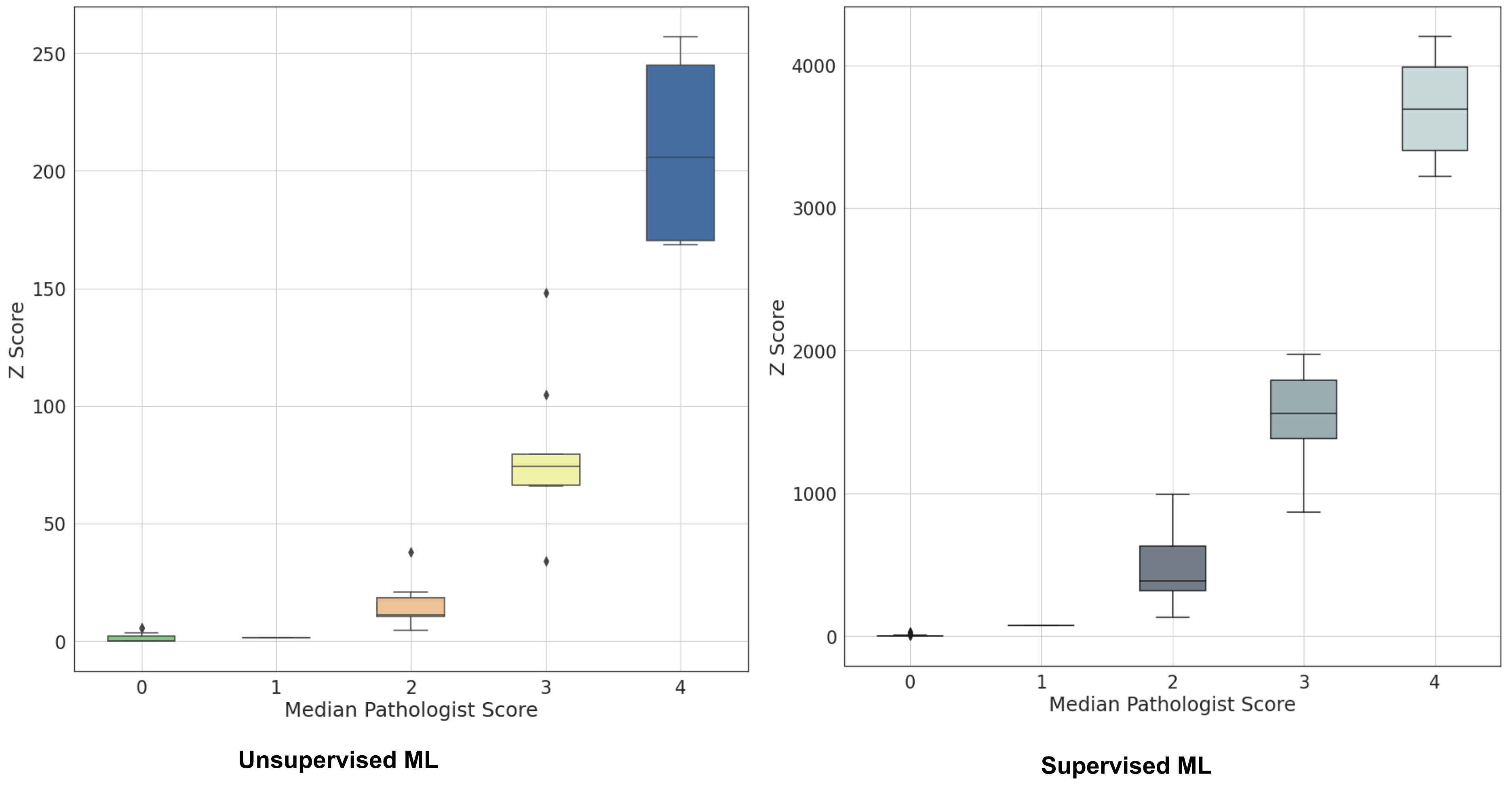

Figure 4. Box and whisker plots comparing both strategies demonstrate a strong correlation between the median severity score assigned by pathologists and the Z-scores produced by ML algorithms.

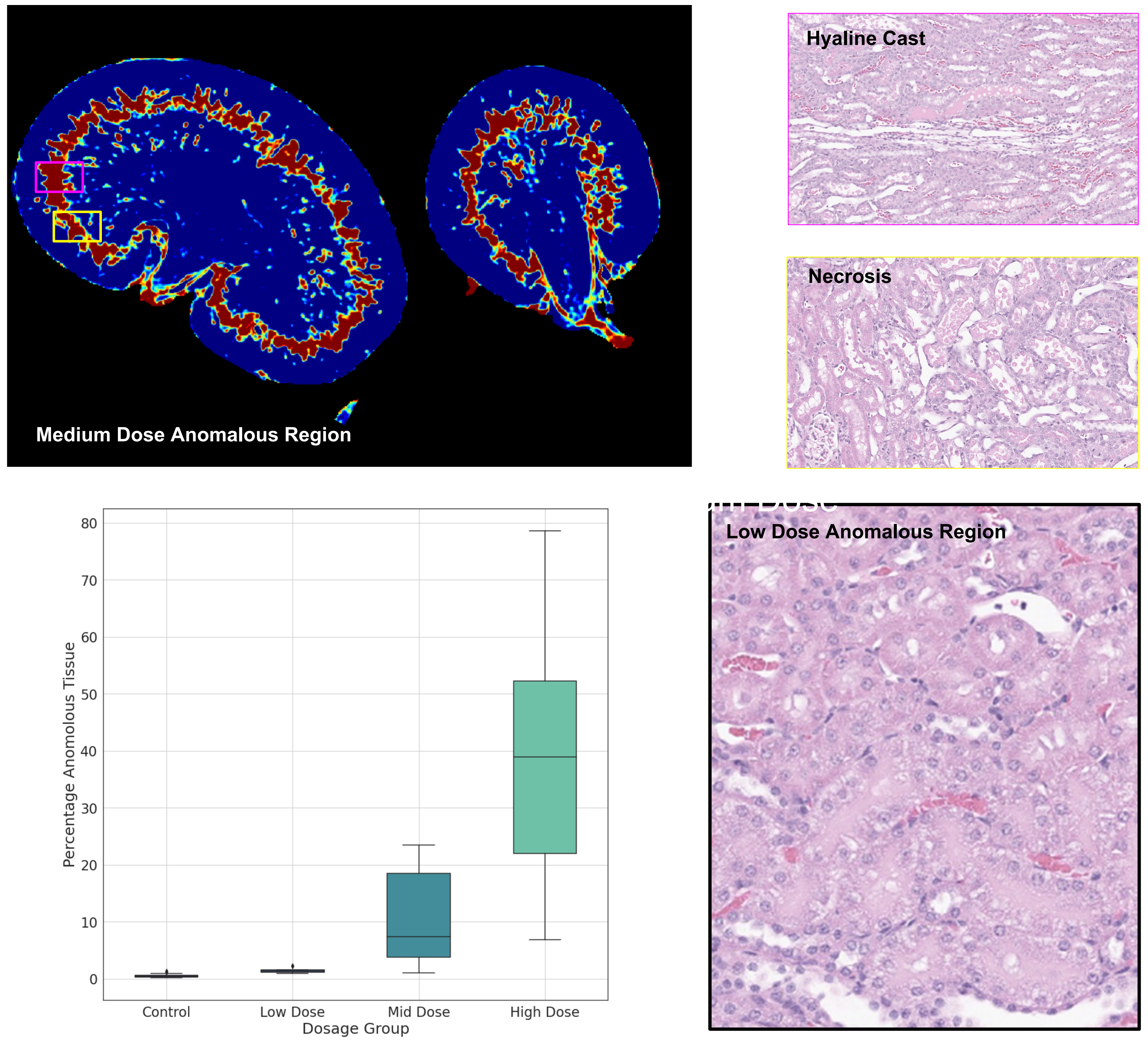

Figure 5. An algorithm trained using a supervised approach identified the anatomic compartment with histopathology findings (upper panels, red annotations). At the low dose, the algorithm identified anomalous regions not considered noteworthy by the pathologists (lower panels). These regions of histomorphologic difference contained fine vacuolations in basolateral cytoplasm of renal tubular epithelial cells.