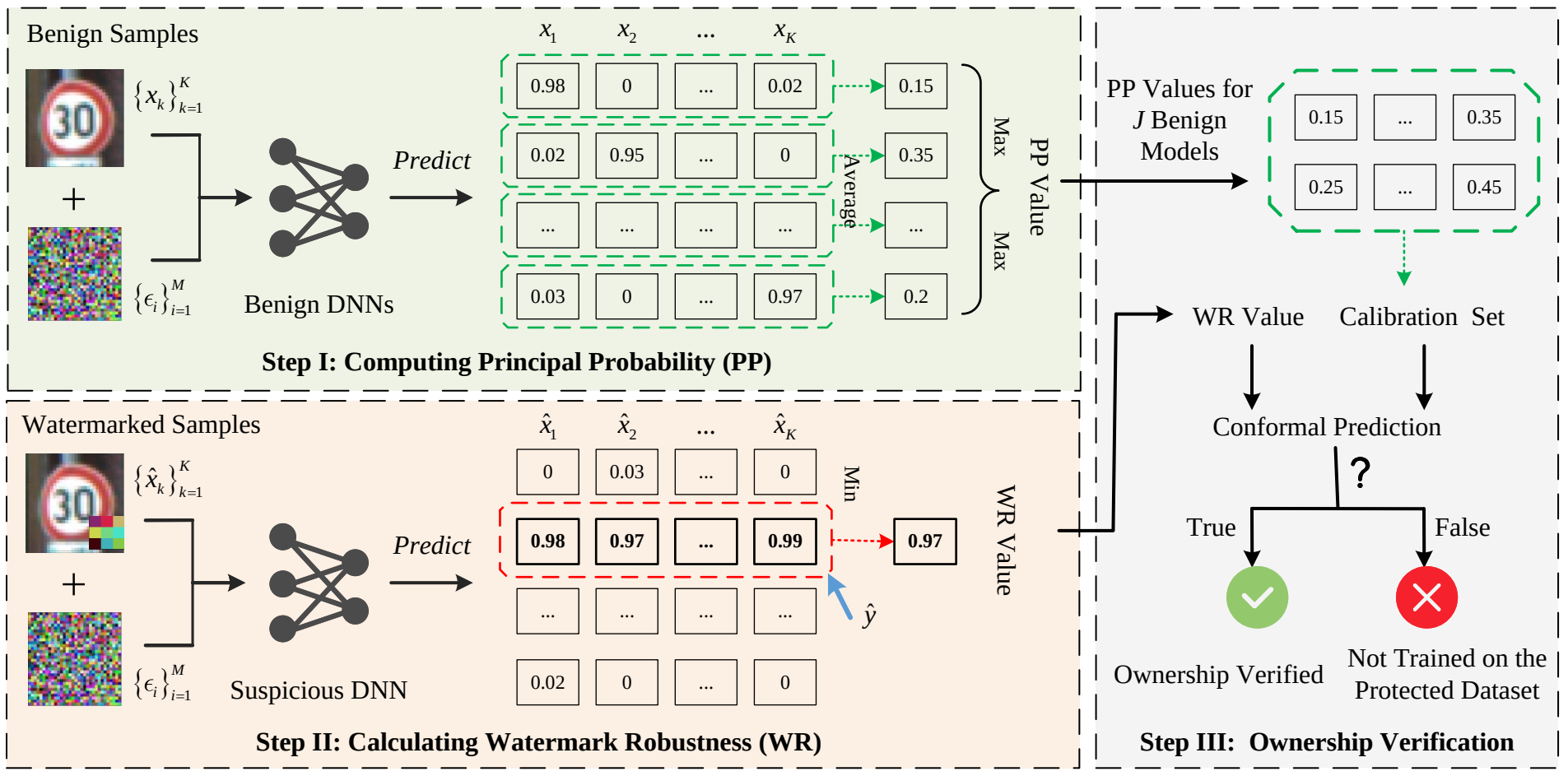
Benign Samples
$\{x_k\}_{k=1}^{K}$
$\{\epsilon_i\}_{i=1}^{M}$
Benign DNNs
Predict
x_1 x_2 ... x_K
0.98 0 ... 0.02
0.02 0.95 ... 0
... ...
0.03 0 ... 0.97
Average
0.15
0.35
...
0.2
Max
Max
PP Value
Step I: Computing Principal Probability (PP)
PP Values for J Benign Models
0.15 ... 0.35
0.25 ... 0.45
WR Value
Calibration Set
Conformal Prediction
?
True
False
Ownership Verified
Not Trained on the Protected Dataset
Step III: Ownership Verification
Watermarked Samples
$\{\hat{x}_k\}_{k=1}^{K}$
$\{\epsilon_i\}_{i=1}^{M}$
Suspicious DNN
Predict
$\hat{x}_1$ $\hat{x}_2$ ... $\hat{x}_K$
0 0.03 ... 0
0.98 0.97 ... 0.99
... ...
0.02 0 ... 0
Min
0.97
$\hat{y}$
WR Value
Step II: Calculating Watermark Robustness (WR)