Species Distribution Modeling of Citizen Science Data as a Classification Problem with Class-conditional Noise: Supplemental Material

Rebecca A. Hutchinson[1,2] **Liqiang He**[1]
Sarah C. Emerson[3]
[1]School of Electrical Engineering and Computer Science
[2]Department of Fisheries and Wildlife
[3]Statistics Department
Oregon State University
Corvallis, OR 97331
{rah,heli,sarah.emerson}@oregonstate.edu

Synthetic data generation process

In the synthetic experiments, all non-intercept coefficients were set to 1, and we varied the intercept coefficients to achieve different class balances and noise levels. For example, when the class balance was low (25%), we have

$$\bar{\psi} = logistic(f(x)) = logistic(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3) = 0.25.$$

Since $\alpha_1 = \alpha_2 = \alpha_3 = 1$, the above equation becomes

$$\bar{\psi} = logistic(\alpha_0 + X_1 + X_2 + X_3) = 0.25.$$

When X_1, X_2 , and X_3 have normal distributions, it is easy to get $\alpha_0 = -1.91$. The other coefficients are given in Table S1. Given the coefficients, we sample the true and observed labels according to the generative model.

For datasets #1-7, we used the logistic link function when generating the data. For dataset #8, we used the probit link. For dataset #9, we scaled the real values of an input vector X to the probability scale around a given target mean X_{targ} using the following process. Firstly, we scale the input vector into range of $[0, 1]$:

$$X_{new} = \frac{X - min(X)}{max(X) - min(X)}.$$

Then we scale the range of X_{new} again to accommodate target mean X_{targ}:

$$X_{new} = 2 \times X_{new} \times min(X_{targ}, 1 - X_{targ}).$$

Finally, we shift the values around the target mean X_{targ}:

$$X_{new} = X_{new} + X_{targ} - mean(X_{new}).$$

N.O.	Name	a_0_h	a_0_m	a_0_l	b_0_h	b_0_m	b_0_l	c_0_h	c_0_m	c_0_l
1	none-cont-logit	-1.645	0	1.645	-0.452	-1.215	-2.381	0.452	-1.215	-2.381
2	none-mix-logit	1.033	-0.5	-2.033	-0.686	-1.423	-2.570	-0.686	-1.423	-2.570
3	none-cat-logit	-0.228	-1.5	-2.272	-0.5	-1.629	-2.747	-0.5	-1.629	-2.747
4	noise-cont-logit	-1.645	0	1.645	-0.452	-1.215	-2.381	0.452	-1.215	-2.381
5	noise-mix-logit	1.033	-0.5	-2.033	-0.686	-1.423	-2.570	-0.686	-1.423	-2.570
6	class-cont-logit	-1.645	0	1.645	-0.452	-1.215	-2.381	0.452	-1.215	-2.381
7	class-mix-logit	1.033	-0.5	-2.033	-0.686	-1.423	-2.570	-0.686	-1.423	-2.570
8	noise-mix-probit	1.033	-0.5	-2.033	-0.686	-1.423	-2.570	-0.686	-1.423	-2.570
9	noise-mix-scale	1.033	-0.5	-2.033	-0.686	-1.423	-2.570	-0.686	-1.423	-2.570

Table S1: Intercept coefficients for the nine different datasets under different class balance and noise levels. "h, m, l" are used to denote three different levels : "high, medium, low". Datasets #1, #4, and #6 have the same coefficients, because the covariates types are the same. Similarly, datasets #2, #5, #7, #8, and #9 have the same coefficients. Dataset # 3 is used to explore the influence of categorical covariates, so its coefficients are different from the others.

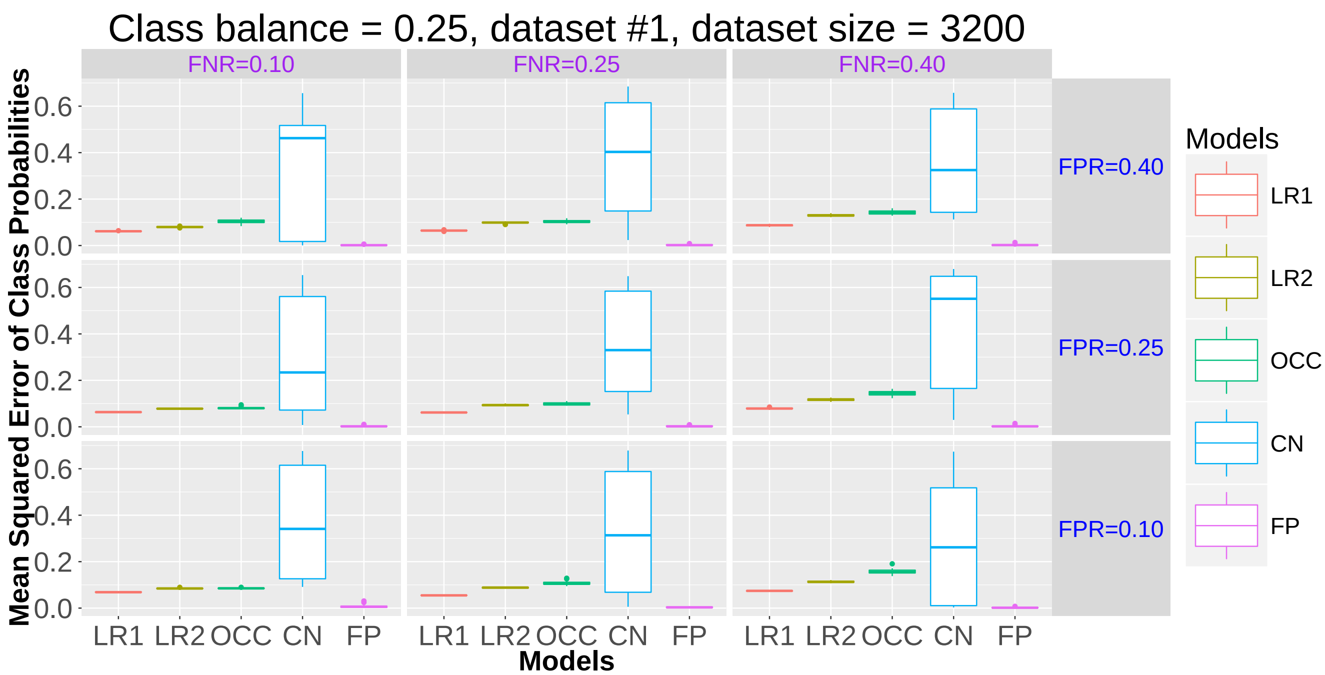

Figure S1: Mean squared error in the class probabilities (ψ) for data-generating model #1 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 25% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

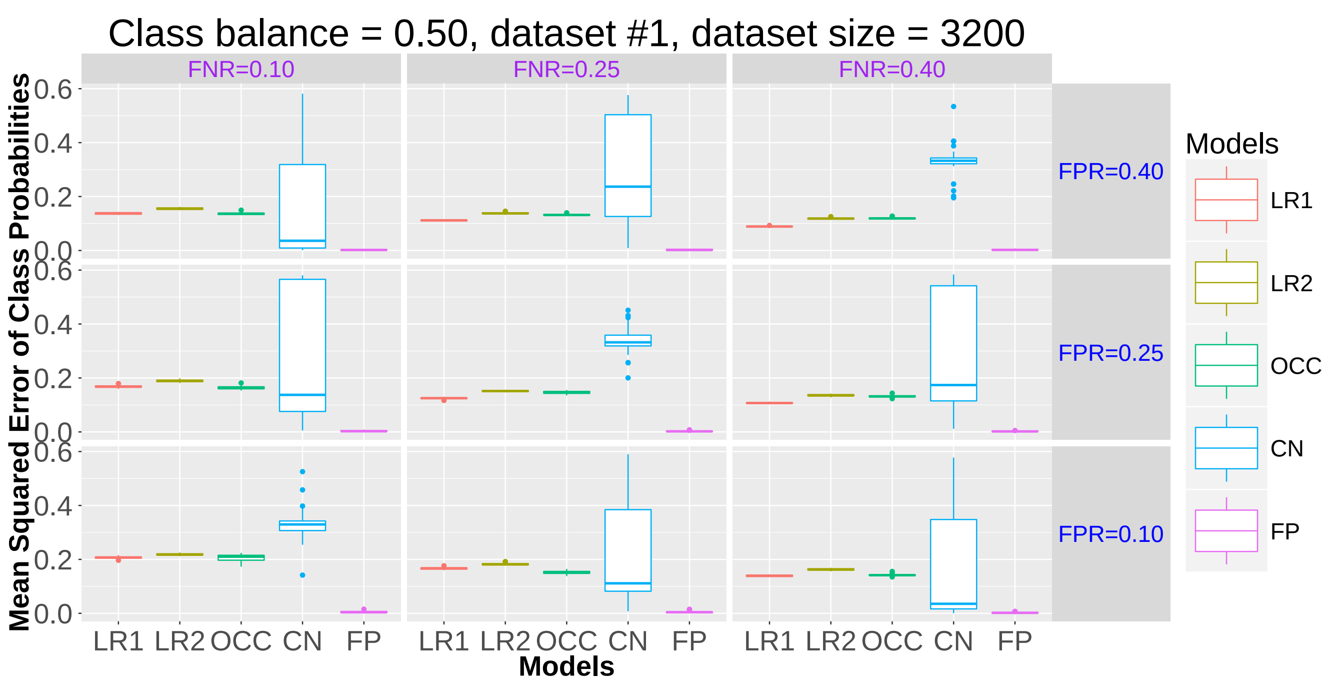

Figure S2: Mean squared error in the class probabilities (ψ) for data-generating model #1 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 50% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets. While the *CN* method often exhibits high variability due to identifiability issues, the perfectly symmetric cases have lower variability.

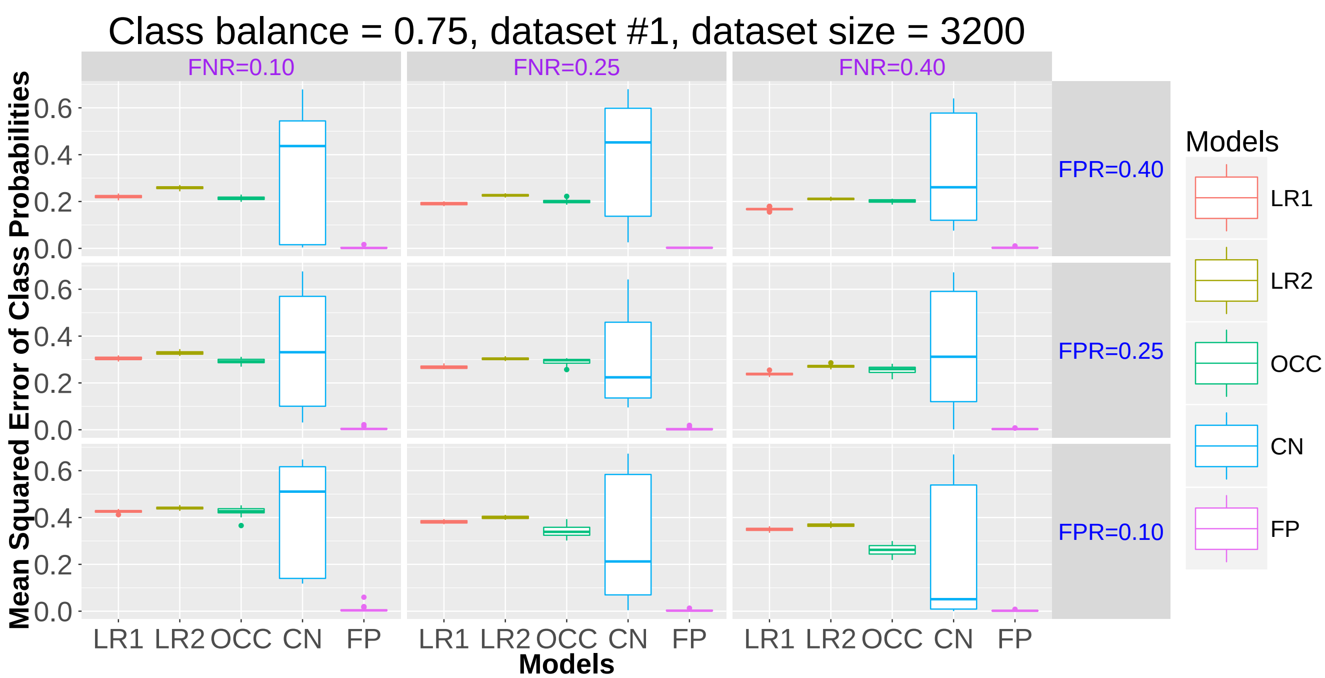

Figure S3: Mean squared error in the class probabilities (ψ) for data-generating model #1 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 75% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

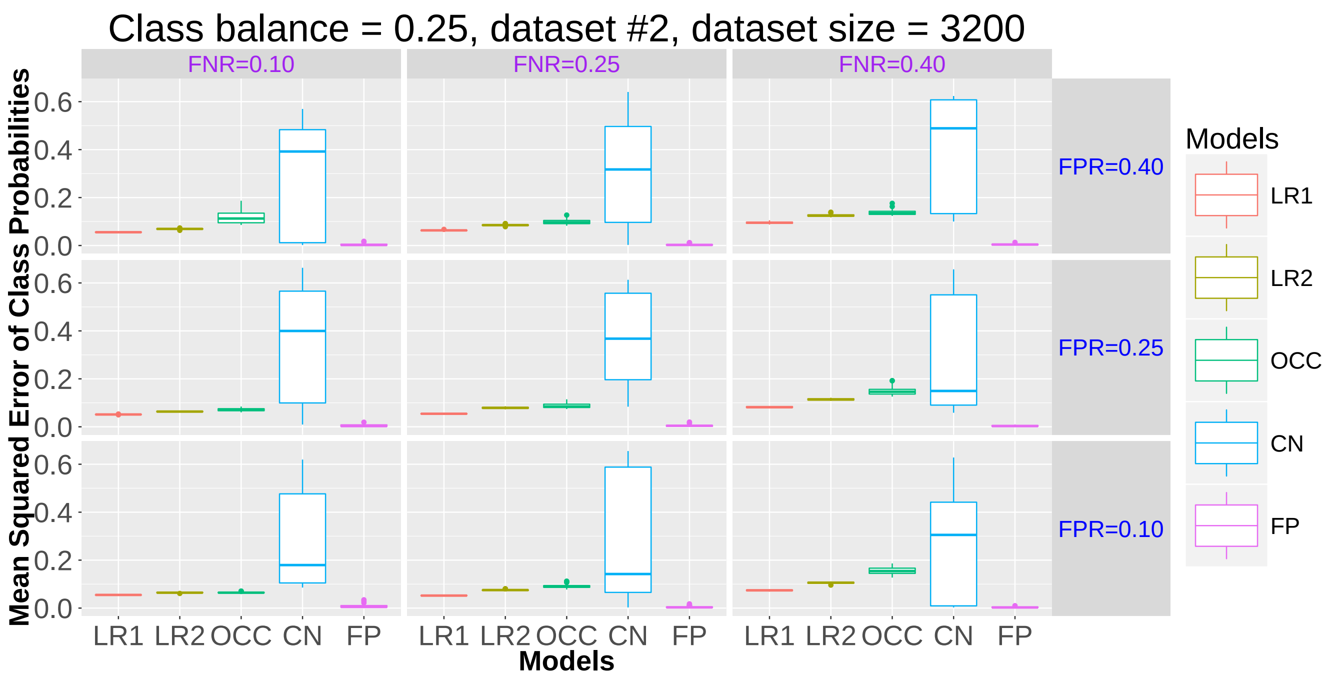

Figure S4: Mean squared error in the class probabilities (ψ) for data-generating model #2 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 25% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

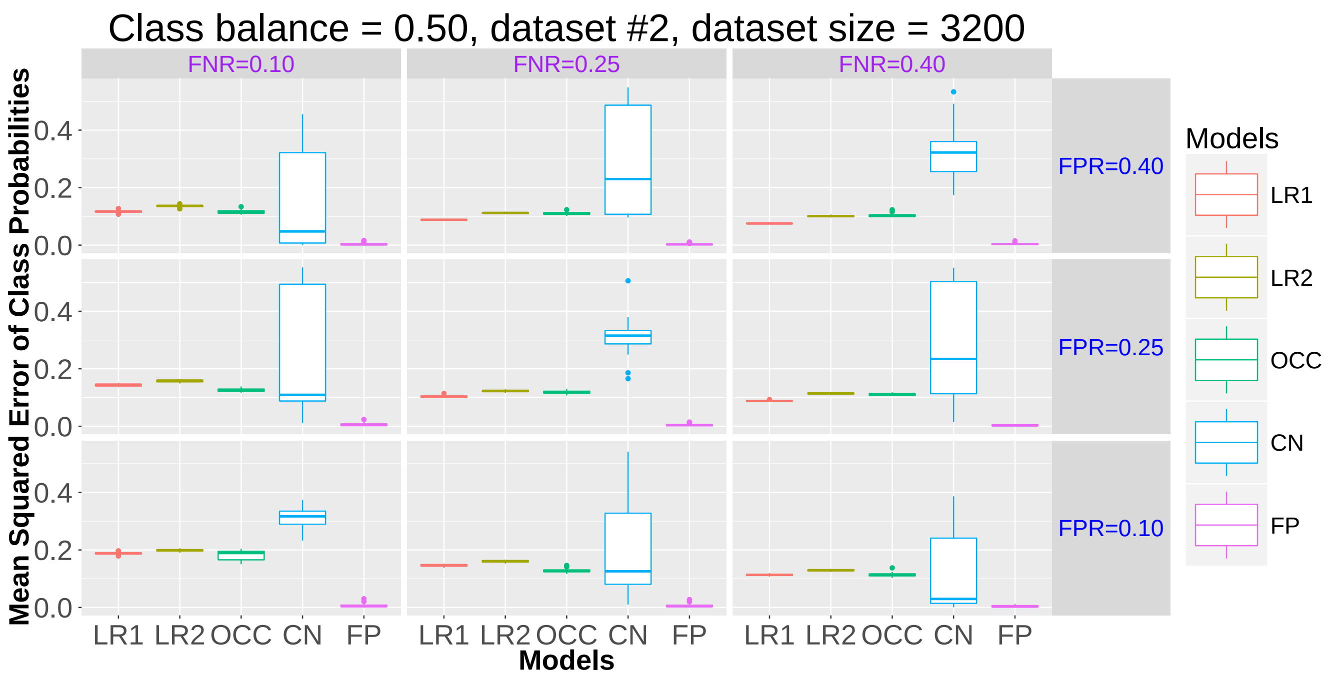

Figure S5: Mean squared error in the class probabilities (ψ) for data-generating model #2 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 50% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets. While the *CN* method often exhibits high variability due to identifiability issues, the perfectly symmetric cases have lower variability.

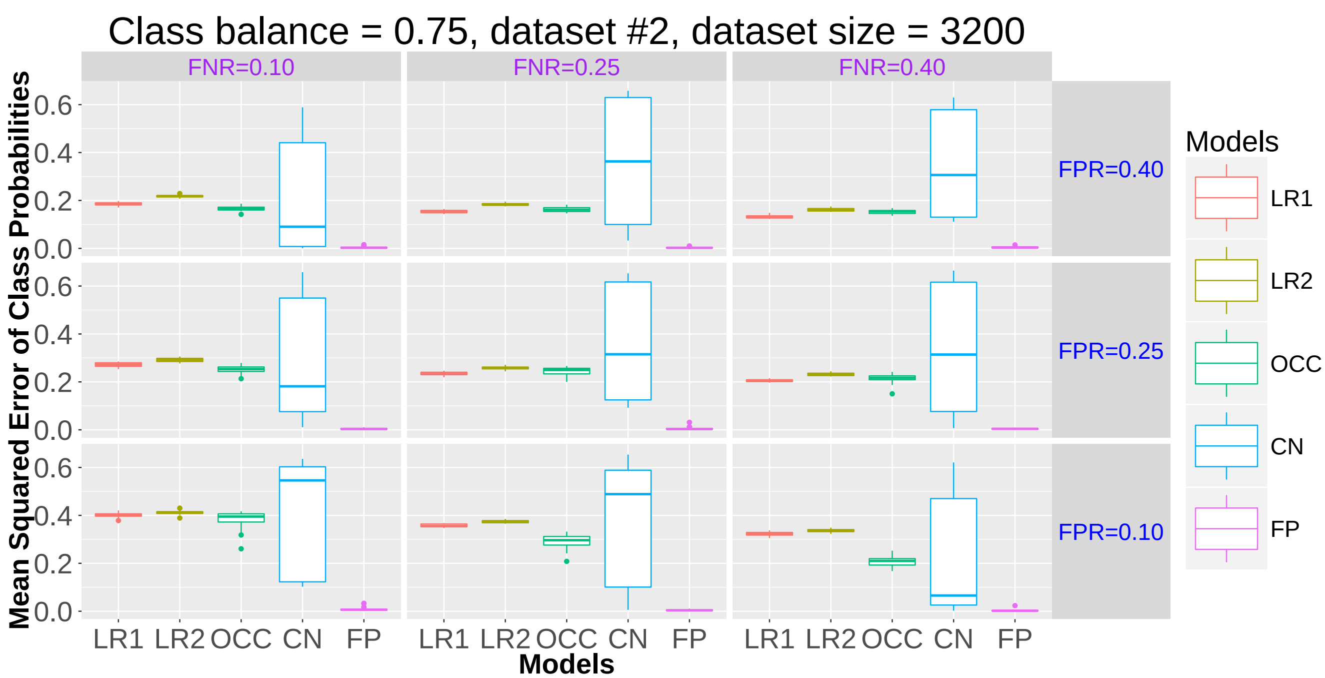

Figure S6: Mean squared error in the class probabilities (ψ) for data-generating model #2 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 75% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

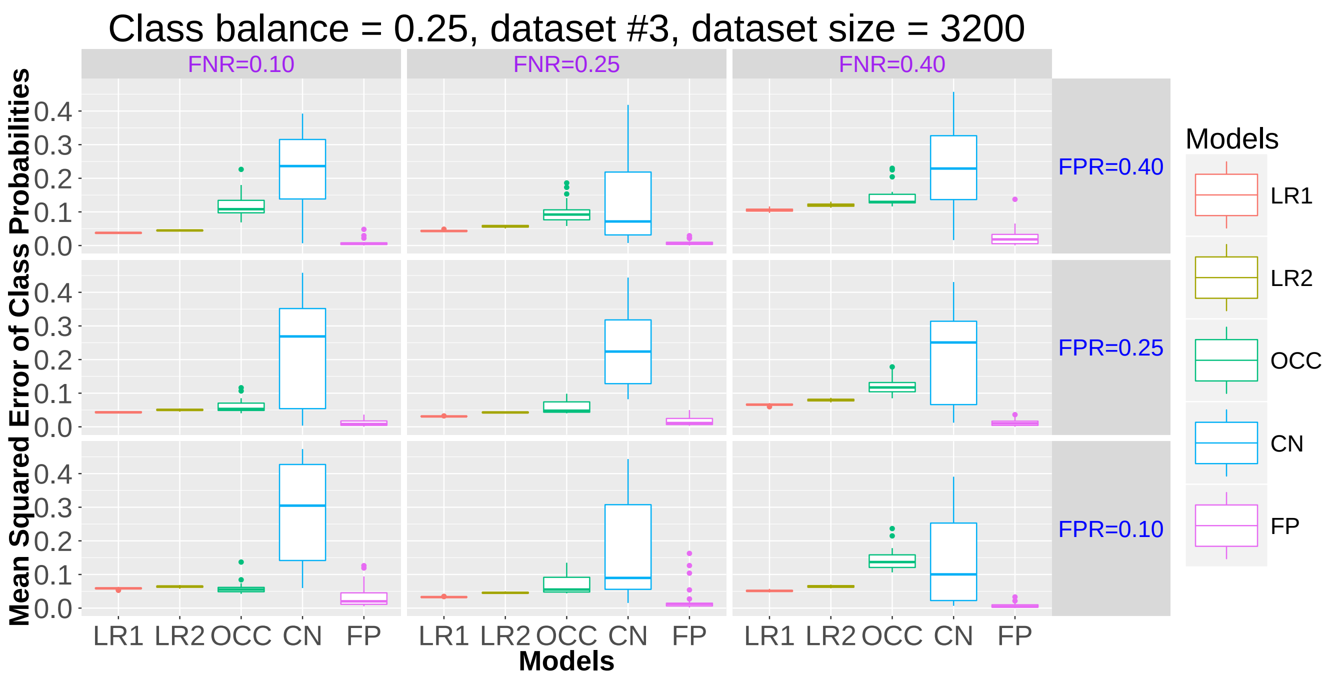

Figure S7: Mean squared error in the class probabilities (ψ) for data-generating model #3 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 25% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

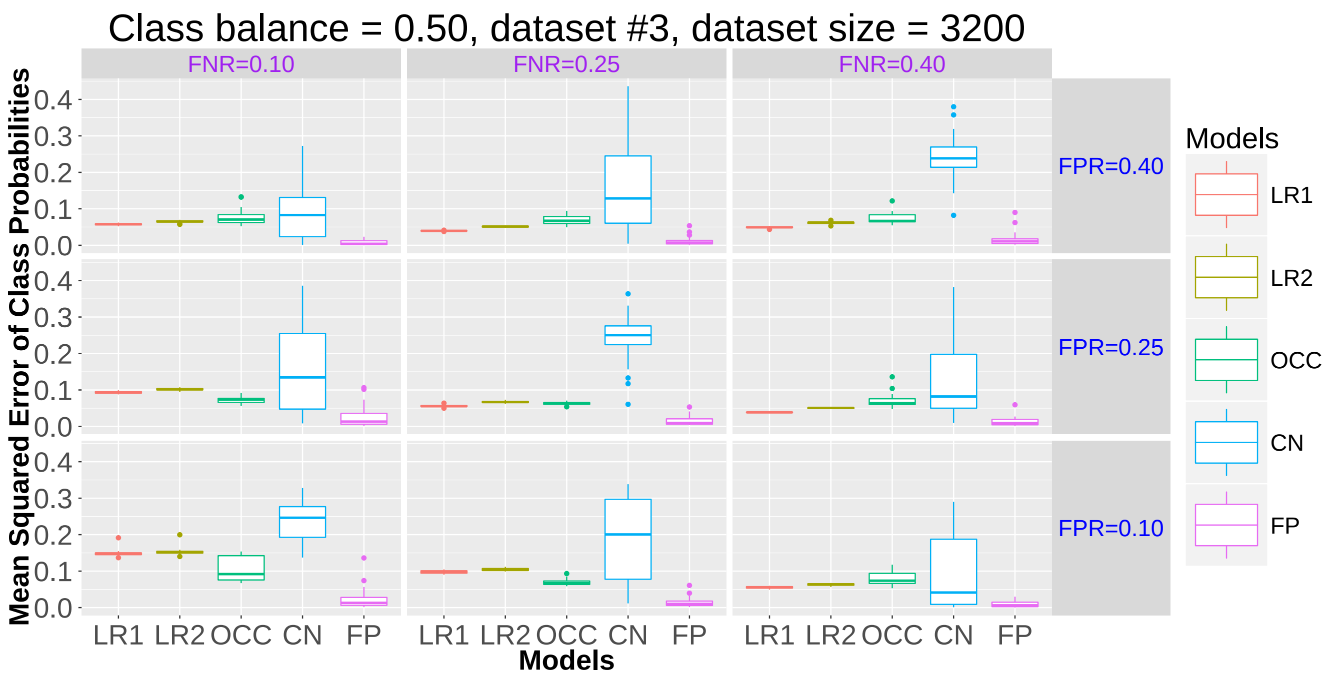

Figure S8: Mean squared error in the class probabilities (ψ) for data-generating model #3 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 50% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets. While the *CN* method often exhibits high variability due to identifiability issues, the perfectly symmetric cases have lower variability.

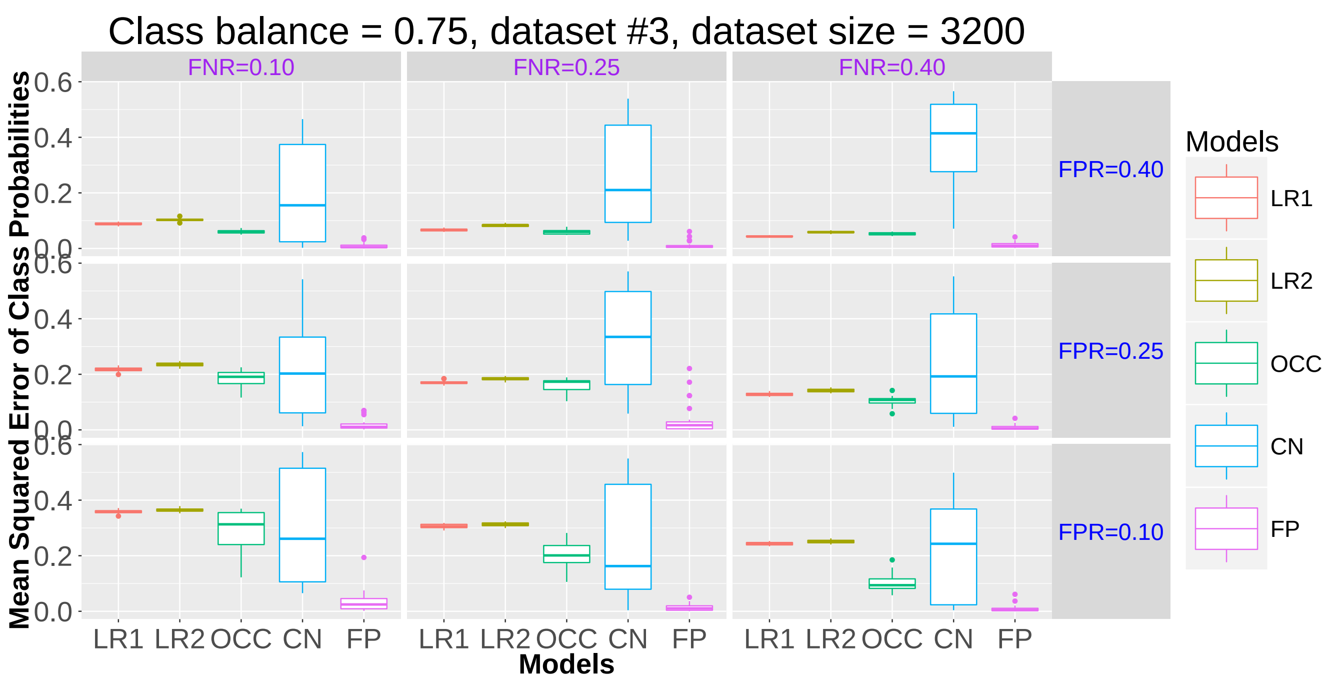

Figure S9: Mean squared error in the class probabilities (ψ) for data-generating model #3 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 75% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

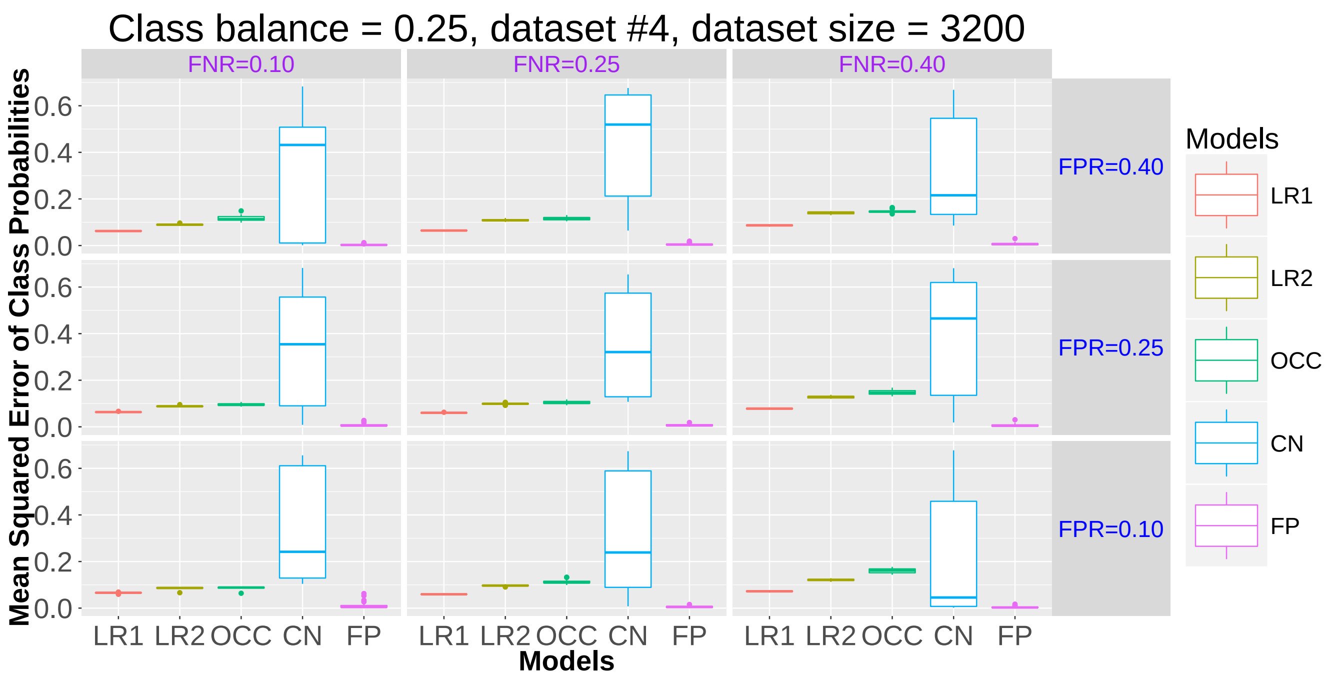

Figure S10: Mean squared error in the class probabilities (ψ) for data-generating model #4 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 25% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

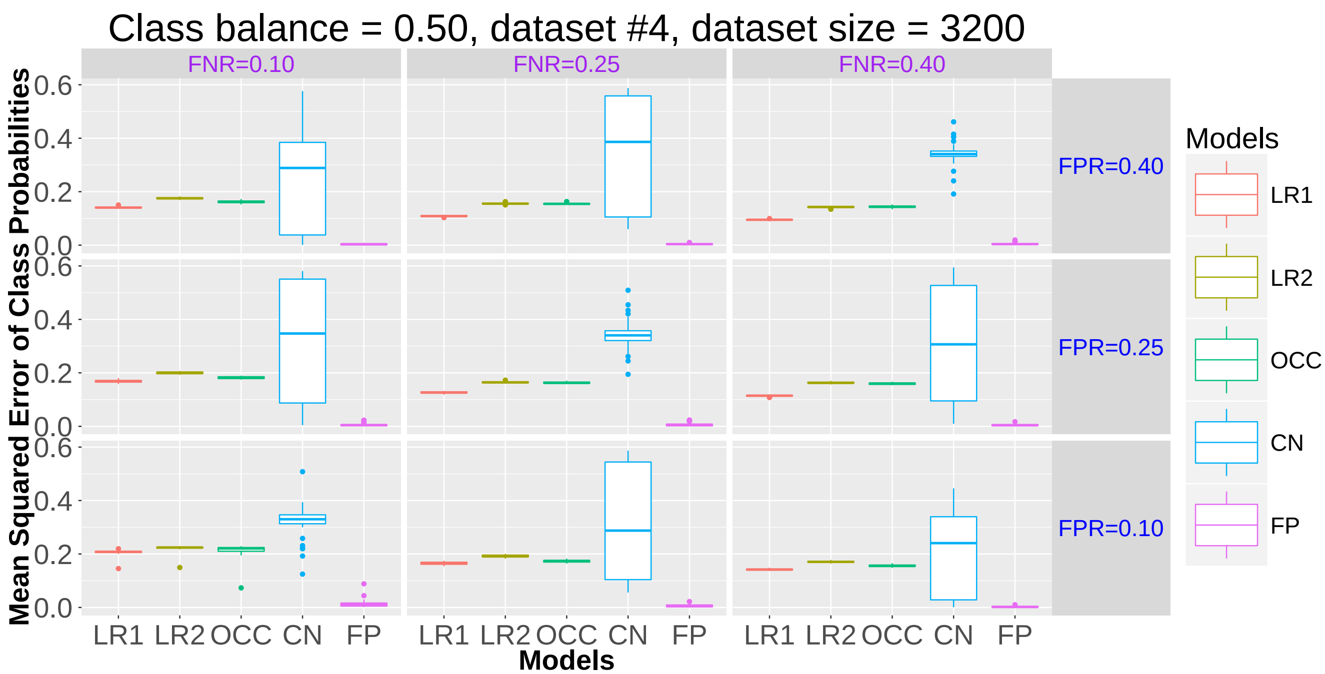

Figure S11: Mean squared error in the class probabilities (ψ) for data-generating model #4 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 50% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets. While the *CN* method often exhibits high variability due to identifiability issues, the perfectly symmetric cases have lower variability.

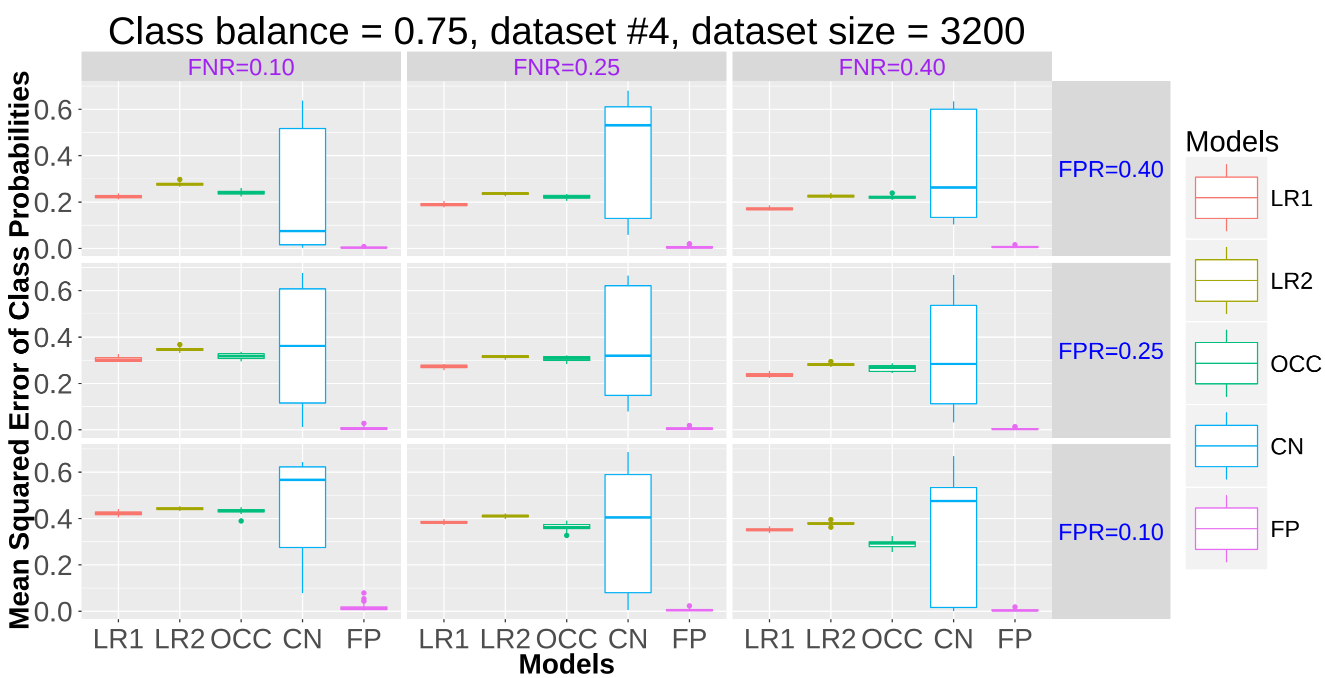

Figure S12: Mean squared error in the class probabilities (ψ) for data-generating model #4 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 75% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

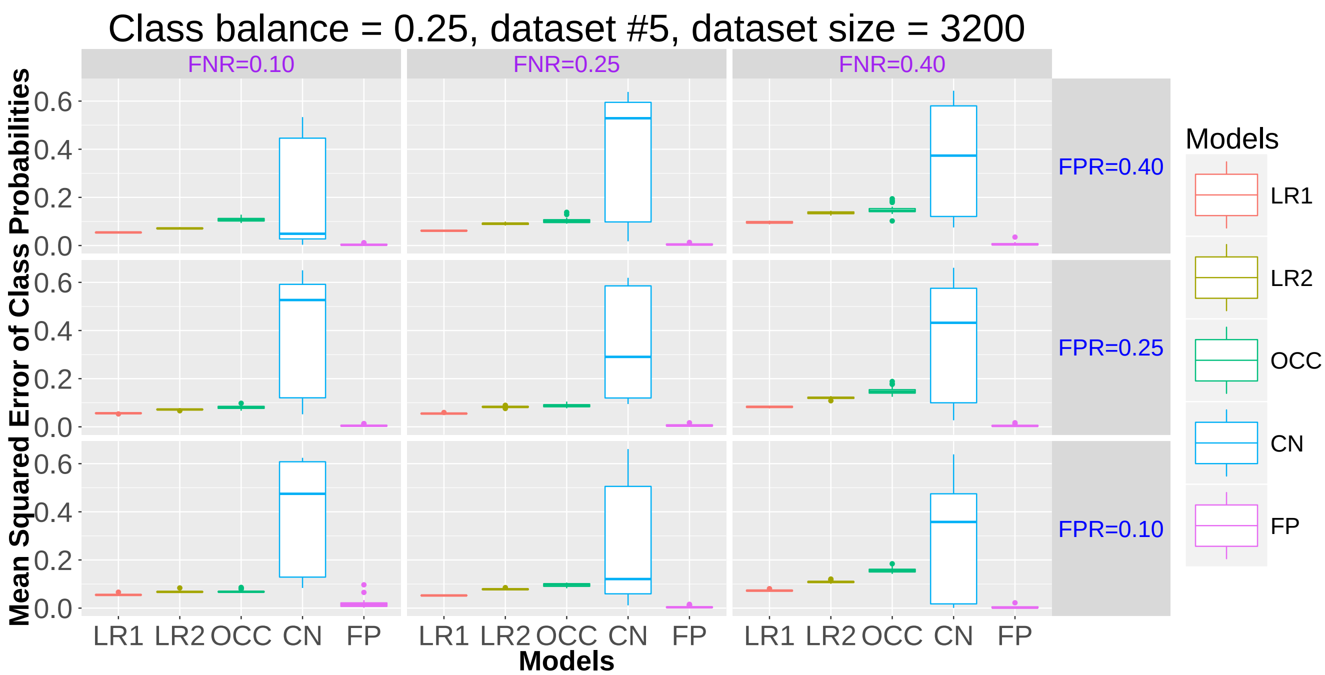

Figure S13: Mean squared error in the class probabilities (ψ) for data-generating model #5 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 25% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

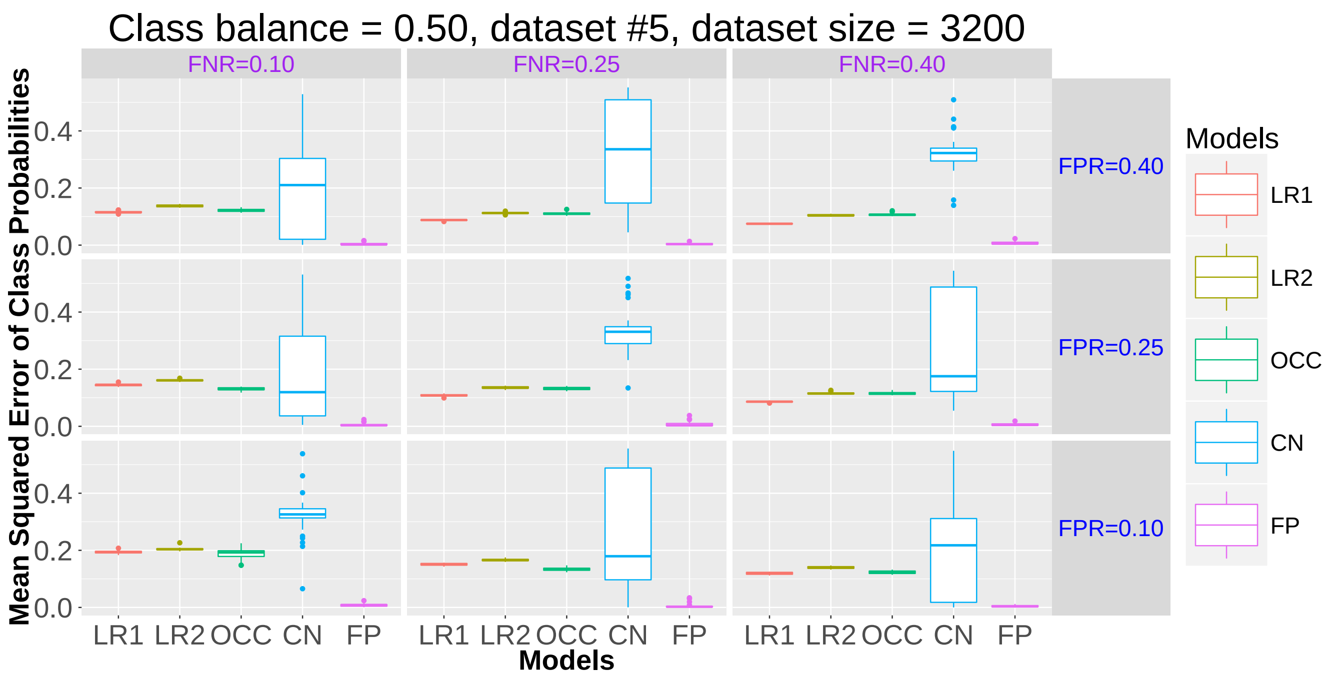

Figure S14: Mean squared error in the class probabilities (ψ) for data-generating model #5 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 50% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets. While the *CN* method often exhibits high variability due to identifiability issues, the perfectly symmetric cases have lower variability.

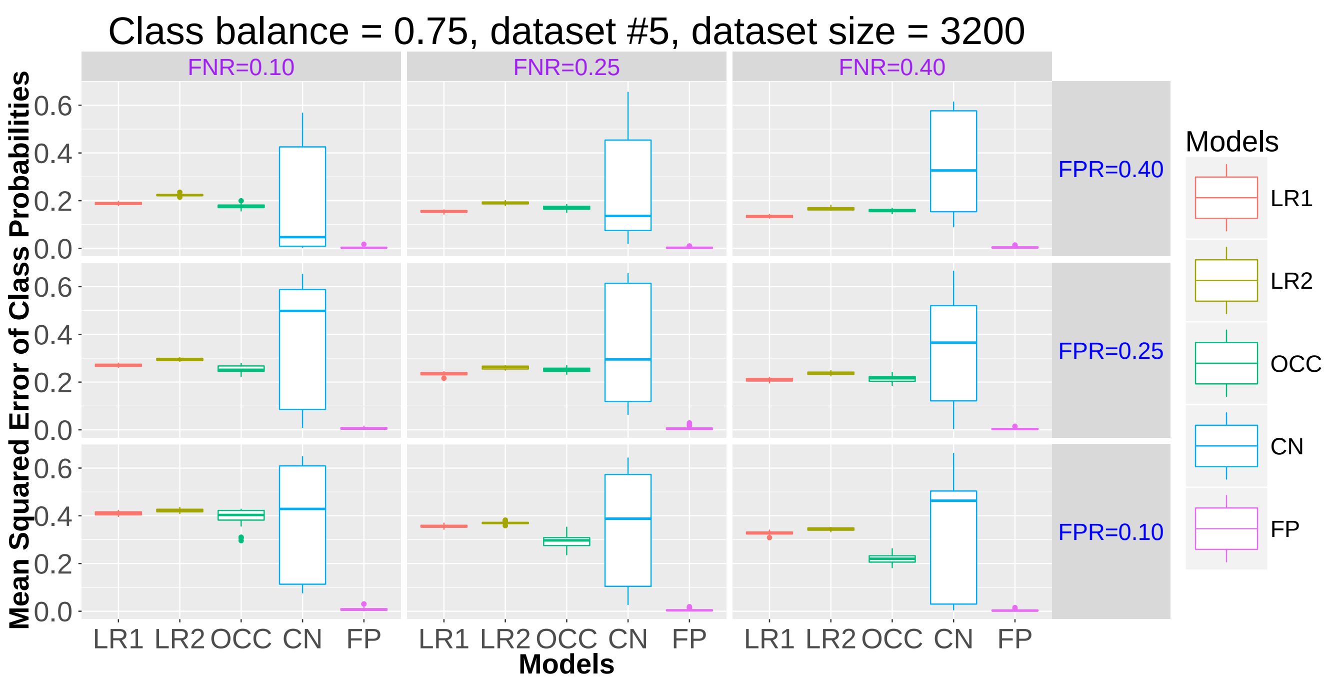

Figure S15: Mean squared error in the class probabilities (ψ) for data-generating model #5 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 75% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

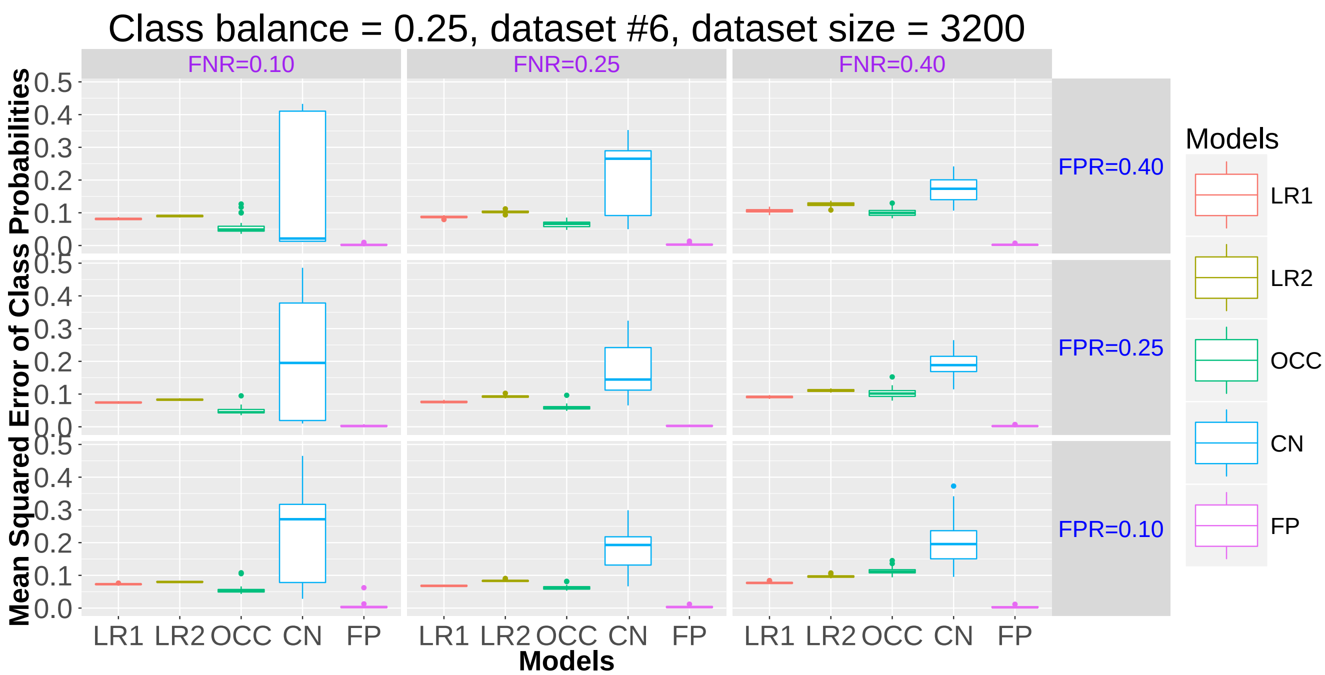

Figure S16: Mean squared error in the class probabilities (ψ) for data-generating model #6 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 25% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

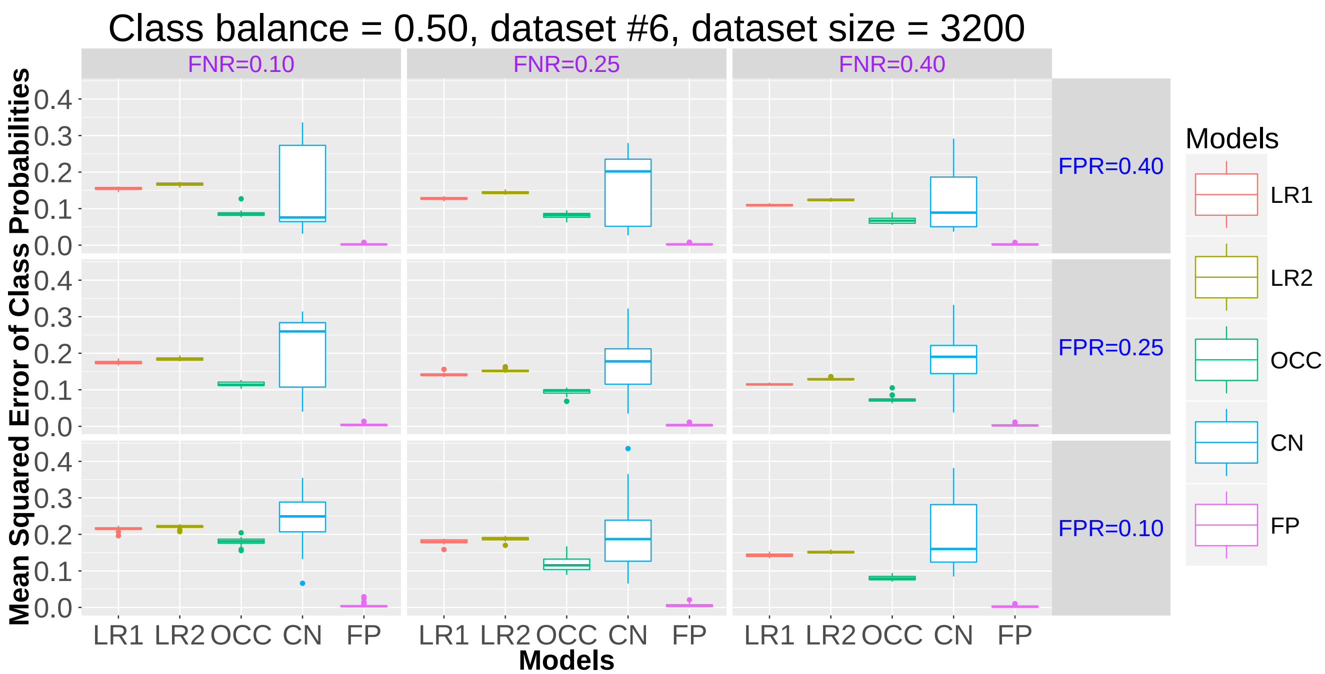

Figure S17: Mean squared error in the class probabilities (ψ) for data-generating model #6 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 50% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets. While the *CN* method often exhibits high variability due to identifiability issues, the perfectly symmetric cases have lower variability.

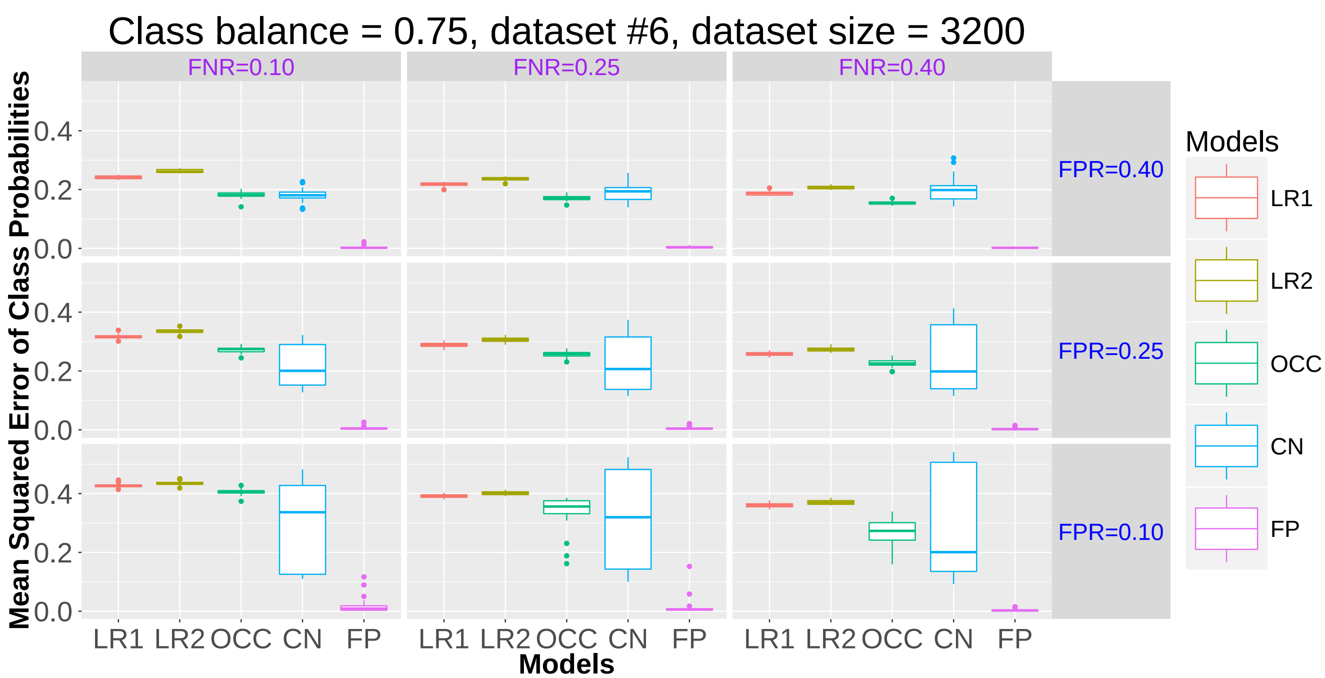

Figure S18: Mean squared error in the class probabilities (ψ) for data-generating model #6 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 75% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

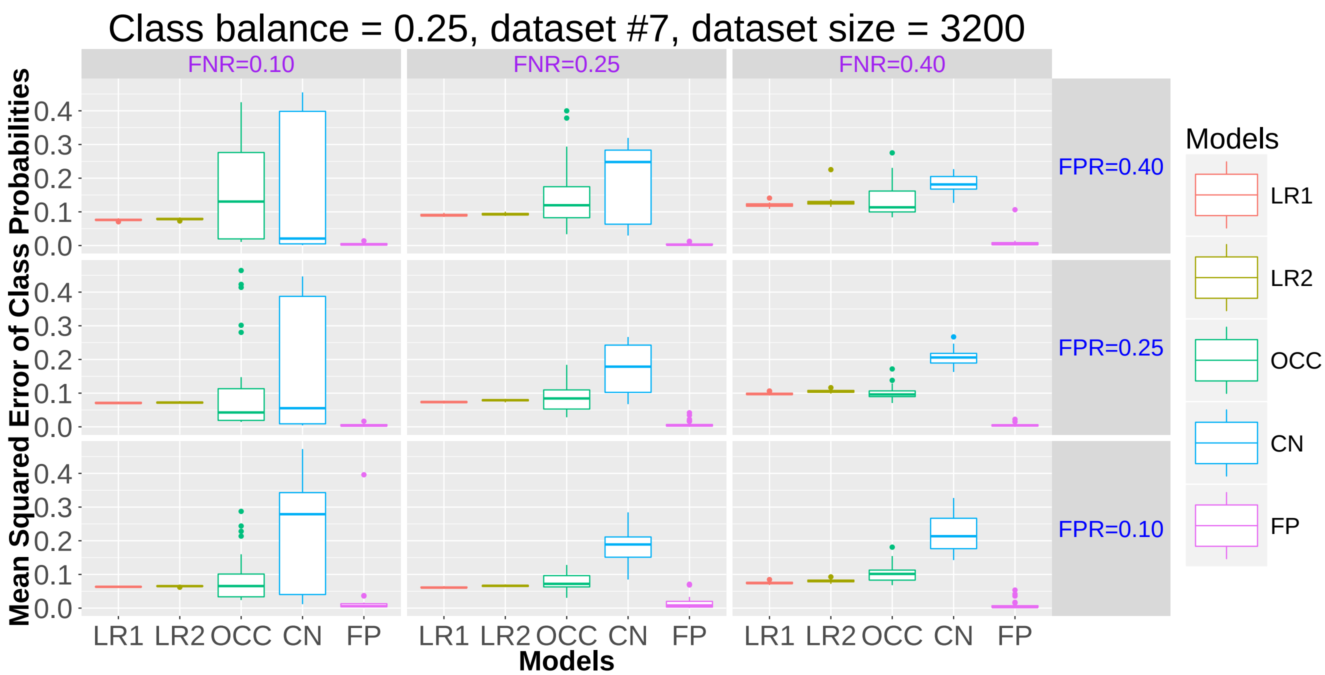

Figure S19: Mean squared error in the class probabilities (ψ) for data-generating model #7 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 25% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

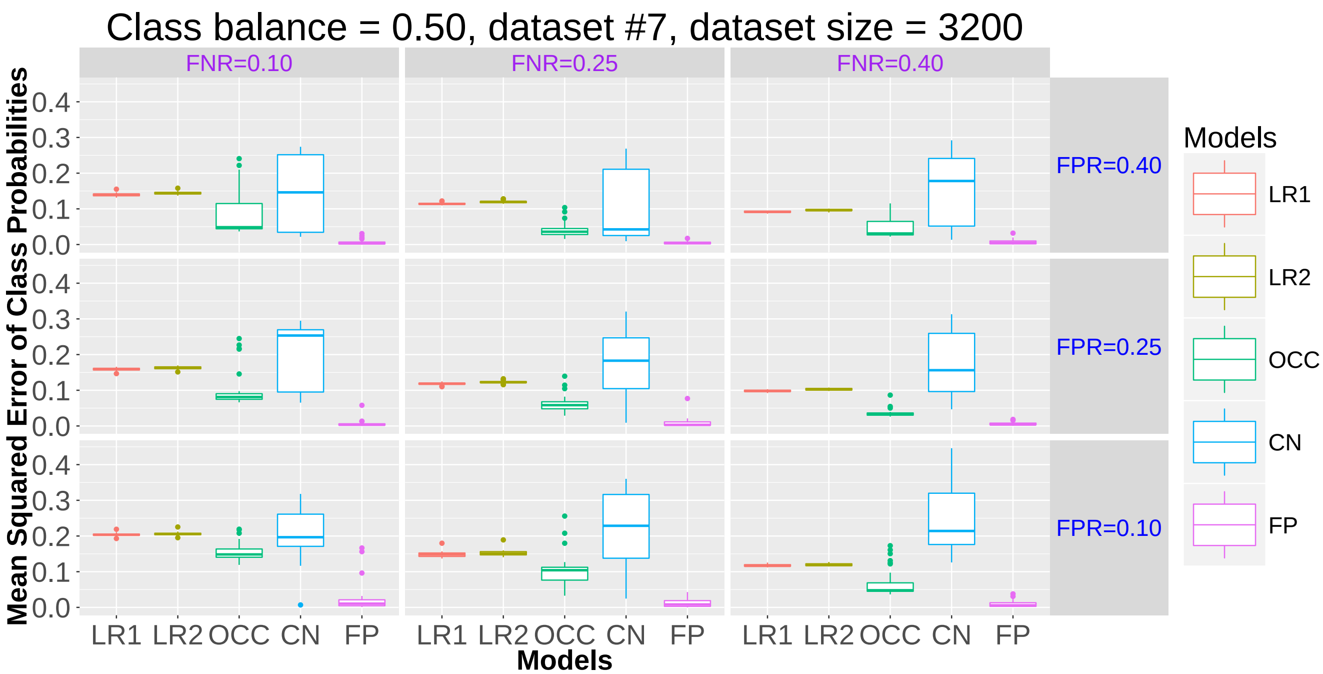

Figure S20: Mean squared error in the class probabilities (ψ) for data-generating model #7 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 50% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets. While the *CN* method often exhibits high variability due to identifiability issues, the perfectly symmetric cases have lower variability.

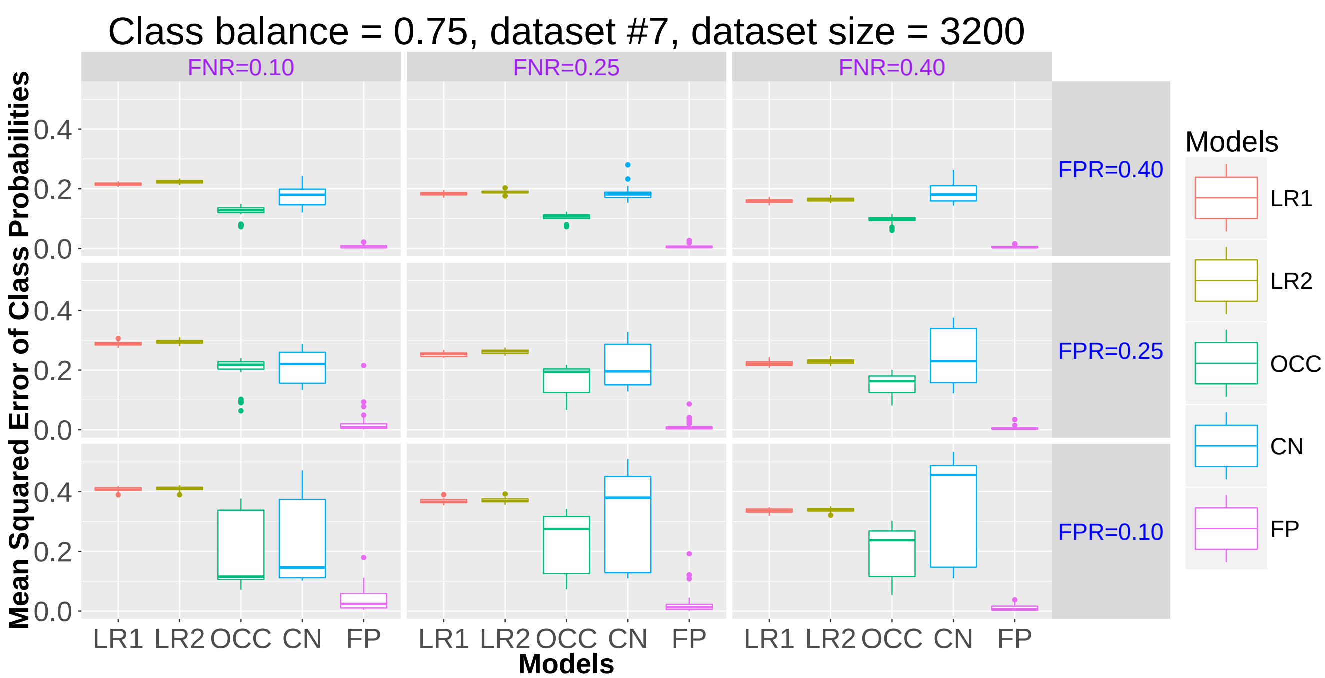

Figure S21: Mean squared error in the class probabilities (ψ) for data-generating model #7 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 75% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

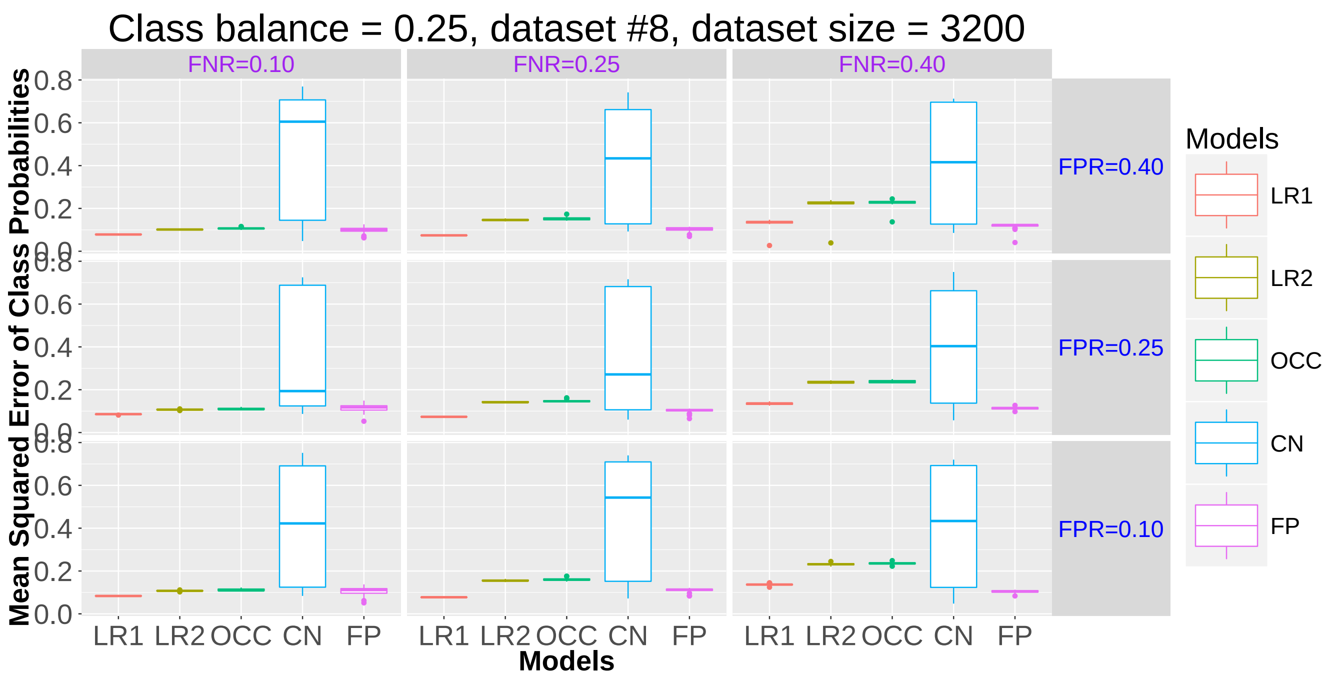

Figure S22: Mean squared error in the class probabilities (ψ) for data-generating model #8 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 25% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

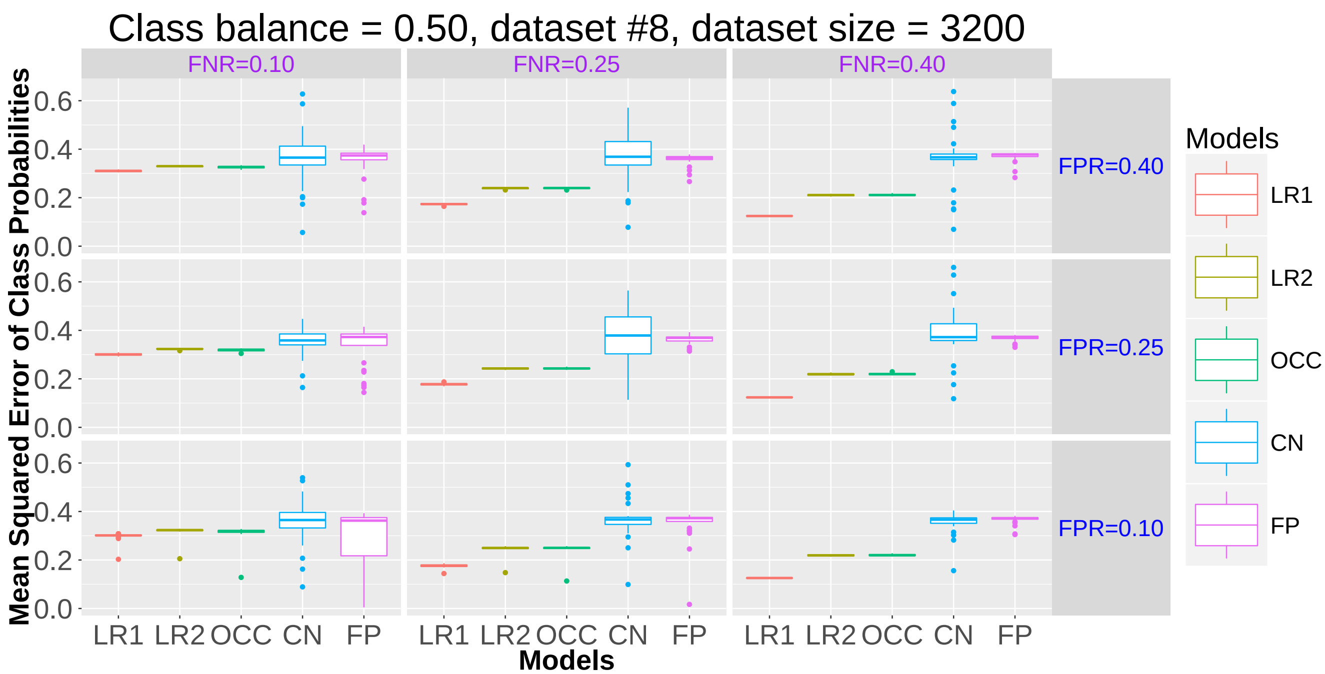

Figure S23: Mean squared error in the class probabilities (ψ) for data-generating model #8 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 50% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets. While the *CN* method often exhibits high variability due to identifiability issues, the perfectly symmetric cases have lower variability.

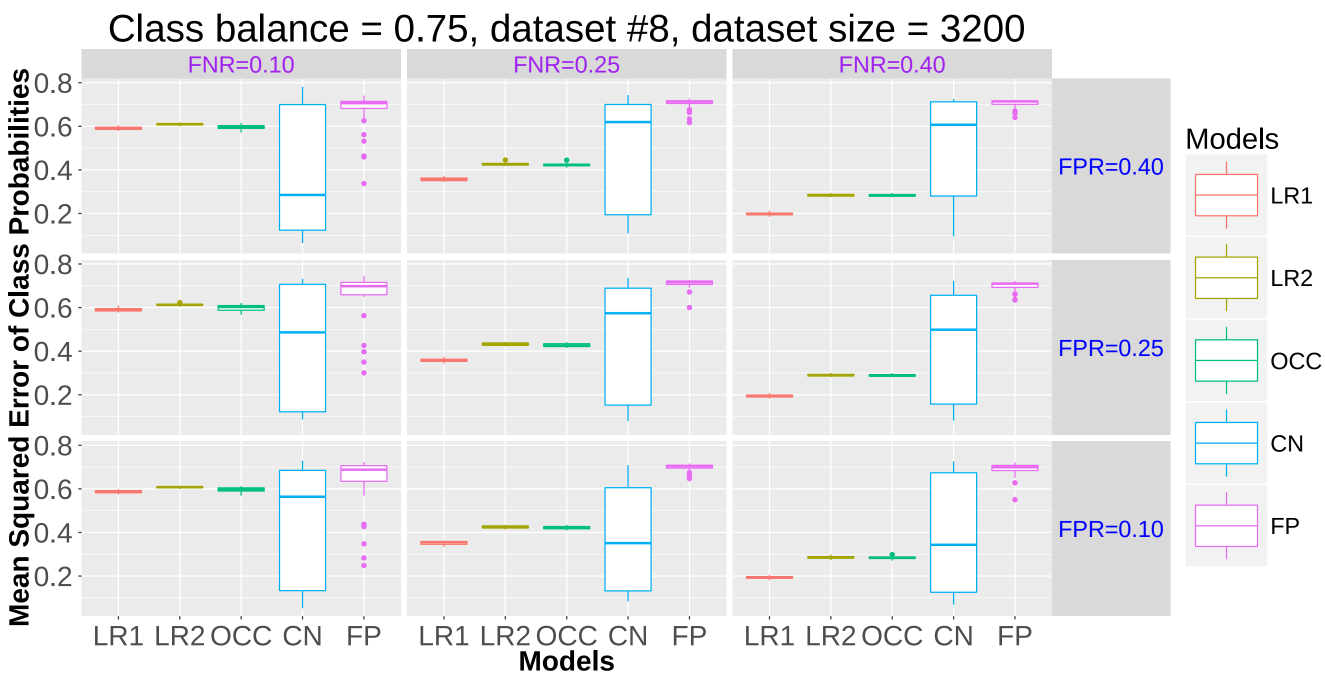

Figure S24: Mean squared error in the class probabilities (ψ) for data-generating model #8 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 75% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

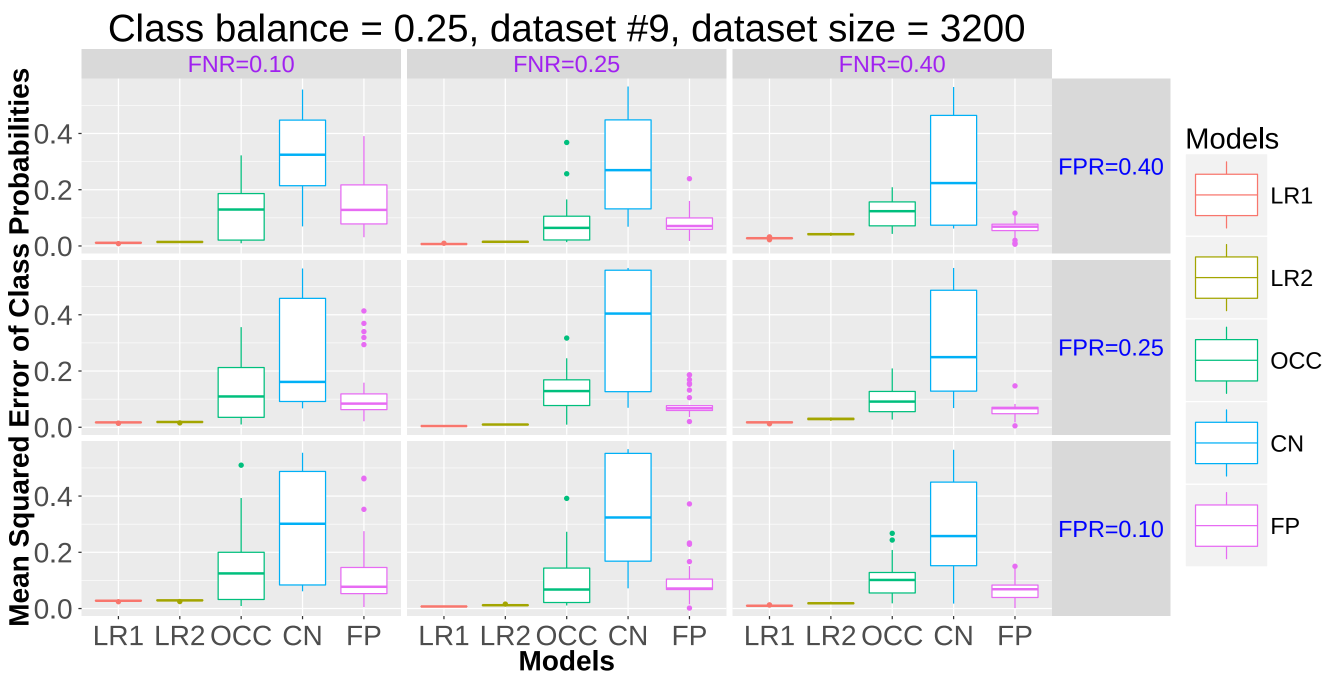

Figure S25: Mean squared error in the class probabilities (ψ) for data-generating model #9 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 25% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

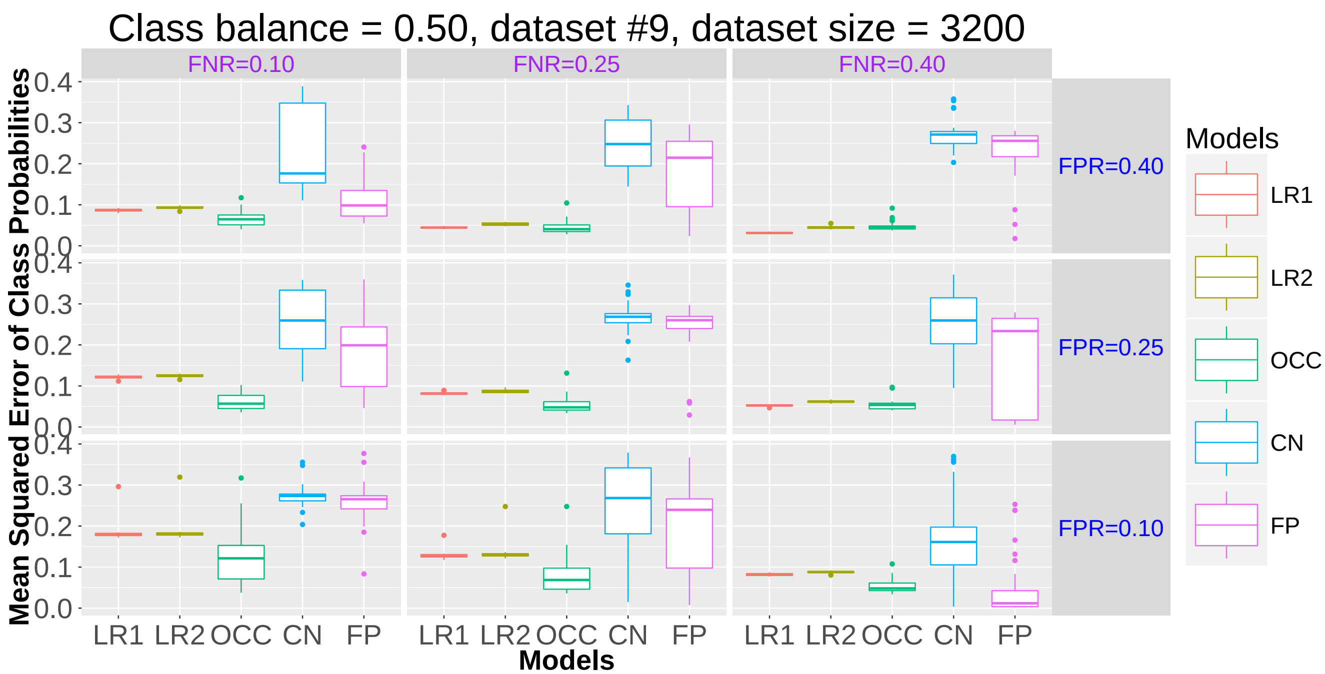

Figure S26: Mean squared error in the class probabilities (ψ) for data-generating model #9 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 50% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets. While the *CN* method often exhibits high variability due to identifiability issues, the perfectly symmetric cases have lower variability.

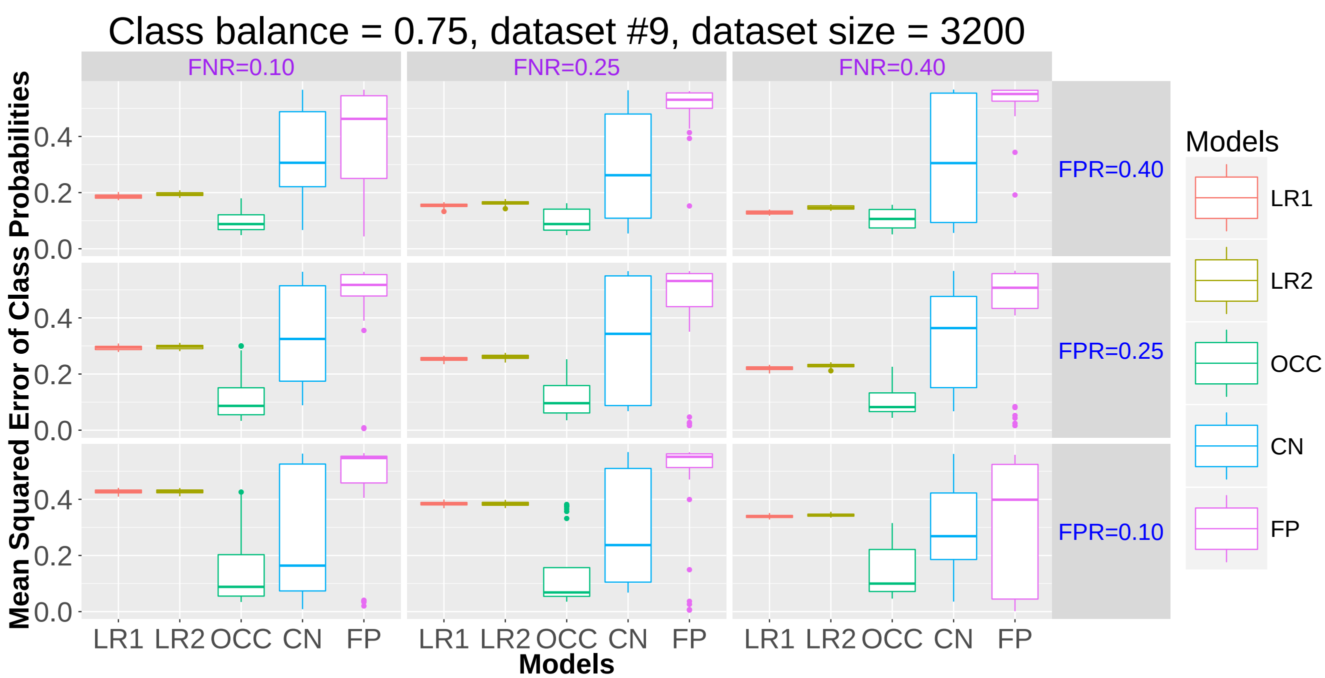

Figure S27: Mean squared error in the class probabilities (ψ) for data-generating model #9 for varying levels of false negative rates (FNR) and false positive rates (FPR) when the true class model had 75% positives. All datasets had 3200 training instances, and each boxplot represents 30 simulated datasets.

Simulated Species	Variables in Submodels	Average rates
1	$\psi = f(ELEVATION, HUMAN.POPULATION)$	0.55
	$\rho = g(DAY)$	0.05
	$\eta = h(EFFORT.HRS, TIME)$	0.2
2	$\psi = f(ELEVATION, HUMAN.POPULATION)$	0.55
	$\rho = g(DAY, TIME)$	0.1
	$\eta = h(EFFORT.HRS, TIME)$	0.4

Table S2: Model forms and average rates of class balance/occupancy, false positives, and false negatives for the species simulated from eBird features.

Model	Variables in Noise Models
1	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$ $\eta = h()$
2	$\rho = g(EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$ $\eta = h(DAY)$
3	$\rho = g(DAY, EFFORT.DIST, N.OBS, TIME)$ $\eta = h(EFFORT.HRS)$
4	$\rho = g(DAY, EFFORT.HRS, N.OBS, TIME)$ $\eta = h(EFFORT.DIST)$
5	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, TIME)$ $\eta = h(N.OBS)$
6	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS)$ $\eta = h(TIME)$
7	$\rho = g(EFFORT.DIST, N.OBS, TIME)$ $\eta = h(DAY, EFFORT.HRS)$
8	$\rho = g(EFFORT.HRS, N.OBS, TIME)$ $\eta = h(DAY, EFFORT.DIST)$
9	$\rho = g(EFFORT.HRS, EFFORT.DIST, TIME)$ $\eta = h(DAY, N.OBS)$
10	$\rho = g(EFFORT.HRS, EFFORT.DIST, N.OBS)$ $\eta = h(DAY, TIME)$
11	$\rho = g(DAY, N.OBS, TIME)$ $\eta = h(EFFORT.HRS, EFFORT.DIST)$
12	$\rho = g(DAY, EFFORT.DIST, TIME)$ $\eta = h(EFFORT.HRS, N.OBS)$
13	$\rho = g(DAY, EFFORT.DIST, N.OBS)$ $\eta = h(EFFORT.HRS, TIME)$
14	$\rho = g(DAY, EFFORT.HRS, TIME)$ $\eta = h(EFFORT.DIST, N.OBS)$
15	$\rho = g(DAY, EFFORT.HRS, N.OBS)$ $\eta = h(EFFORT.DIST, TIME)$
16	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST)$ $\eta = h(N.OBS, TIME)$
17	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$ $\eta = h(EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$
18	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$ $\eta = h(DAY, EFFORT.HRS, EFFORT.DIST, TIME)$
19	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$ $\eta = h(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS)$
20	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$ $\eta = h(DAY, EFFORT.DIST, N.OBS, TIME)$
21	$\rho = g(DAY, EFFORT.HRS, EFFORT.DIST, N.OBS, TIME)$ $\eta = h(DAY, EFFORT.HRS, N.OBS, TIME)$

Table S3: Models considered for the eBird species. Models 1-16 assign each of the five noise features to exactly one of the two submodels. Models 17-21 assign all five noise features to one submodel and all except one feature to the other submodel.

California						
Common Name	Scientific Name	Report frequency	Model selected	Est. avg. occ. prob.	Est. avg. false neg. prob.	Est. avg. false pos. prob.
American crow	*Corvus brachyrhynchos*	0.34	12	0.52	0.40	0.037
Song sparrow	*Melospiza melodia*	0.32	7	0.52	0.37	0.013
Red-winged blackbird	*Agelaius phoeniceus*	0.23	17	0.32	0.28	0.029
Nuttall's woodpecker	*Picoides nuttallii*	0.18	11	0.57	0.42	0.0058
Western kingbird	*Tyrannus verticalis*	0.16	18	0.38	0.36	0.031
Western wood pewee	*Contopus sordidulus*	0.13	13	0.43	0.39	0.038
Northern rough-winged swallow	*Stelgidopteryx serripennis*	0.13	9	0.46	0.50	0.016
Sim1	*Simulus primus*	0.48	7	0.56	0.17	0.054
Sim2	*Simulus secondus*	0.38	20	0.53	0.40	0.099

New York						
Common Name	Scientific Name	Report frequency	Model selected	Est. avg. occ. prob.	Est. avg. false neg. prob.	Est. avg. false pos. prob.
Red-winged blackbird	*Agelaius phoeniceus*	0.59	18	0.31	0.30	0.077
Song sparrow	*Melospiza melodia*	0.56	5	0.50	0.15	0.021
American crow	*Corvus brachyrhynchos*	0.49	17	0.72	0.28	0.035
Red-eyed vireo	*Vireo olivaceus*	0.26	20	0.28	0.22	0.095
Wood thrush	*Hylocichla mustelina*	0.21	5	0.42	0.46	0.012
Eastern wood pewee	*Contopus virens*	0.14	2	0.40	0.32	0.051
Indigo bunting	*Passerina cyanea*	0.14	1	0.30	0.0	0.061
Veery	*Catharus fuscescens*	0.13	13	0.47	0.30	0.024
Sim1	*Simulus primus*	0.48	15	0.62	0.20	0.048
Sim2	*Simulus secondus*	0.40	20	0.54	0.40	0.11

Table S4: Species modeled with the eBird Reference Dataset. The table also indicates the overall freqency of positive reports of the species in the data, the model selected for the *FP* method, and the estimated average occupancy (class balance), false negative, and false positive rates. Note that these results deserve further evaluation from an ecological perspecitve; for example, a false negative rate of 0 for the Indigo bunting may be a sign of model overfitting and not a realistic estimate.

Model	Validation NLL	Test NLL	Test MSE on ψ	Test MSE on η	Test MSE on ρ
1	2406.6	2278.1	0.055	0.00042	0.0076
2	2375.1	**2039.6**	**0.023**	0.000079	**0.000073**
3	2449.1	2438.1	0.078	0.018	0.0078
4	2398.8	2281.3	0.056	0.00037	0.0076
5	2391.0	2264.0	0.053	0.00036	0.0076
6	2602.6	2264.2	0.043	0.030	0.00074
7	**2238.6**	2065.4	0.023	0.018	0.00023
8	2394.6	2046.8	0.024	0.000046	0.00011
9	2383.1	2045.3	0.024	0.000054	0.000098
10	2478.0	2215.4	0.034	0.038	0.0022
11	2432.3	2300.1	0.048	0.039	0.0025
12	2441.3	2409.8	0.072	0.019	0.0078
13	2404.0	2053.0	0.025	**0.000021**	0.00014
14	2383.6	2268.8	0.055	0.00031	0.0076
15	2329.6	2089.9	0.026	0.019	0.00025
16	2430.5	2154.0	0.028	0.028	0.00070
17	2458.0	2071.2	0.027	0.00012	0.00031
18	2345.7	2091.7	0.030	0.00011	0.00032
19	2396.1	2058.0	0.025	0.00011	0.00018
20	2445.0	2079.7	0.028	0.00014	0.00032
21	2425.3	2064.3	0.026	0.00077	0.00030
OCC	-	-	0.093	0.00089	-
CN	-	-	0.070	0.086	0.0078

Table S5: Model selection results for *Sim1* in CA. Bold-numbered models (or their symmetric analogs) are consistent with the true data-generating model. Bold values indicate the best value in each column. Here, model 7 is chosen using the validation set even though it is not consistent with the data-generating model. On the test set, model 2 performs best on the class model, but model 7 is nearly tied with it. All 21 *FP* models outperform the *OCC* model, and all but two outperform the *CN* model on ψ.

Model	Validation NLL	Test NLL	Test MSE on ψ	Test MSE on η	Test MSE on ρ
1	2174.7	2170.9	0.018	0.0028	0.034
2	2124.3	2218.2	0.048	0.00070	0.019
3	2400.7	2452.7	0.090	0.044	0.036
4	2179.4	2169.2	**0.017**	0.0028	0.034
5	2178.6	2173.0	0.019	0.0028	0.034
6	2549.5	2821.9	0.16	0.061	0.020
7	2338.7	2545.4	0.10	0.040	0.019
8	2128.4	2223.8	0.050	0.00067	0.019
9	2121.8	2204.2	0.045	0.00064	0.019
10	2984.8	3146.4	0.12	0.084	0.0010
11	2979.9	3143.8	0.12	0.085	0.0013
12	2413.4	2467.1	0.086	0.046	0.036
13	2125.6	2209.5	0.046	0.00060	0.019
14	2181.9	2170.6	0.018	0.0027	0.034
15	2371.7	2639.1	0.12	0.042	0.019
16	2514.8	2705.0	0.14	0.057	0.020
17	2106.4	2119.2	0.024	**0.00027**	0.000096
18	2100.3	2117.4	0.023	0.00043	0.000088
19	2157.4	2261.5	0.059	0.0010	0.018
20	**2099.4**	**2116.5**	0.023	0.00044	0.00010
21	2100.3	2116.7	0.023	0.00043	**0.000064**
OCC	-	-	0.15	0.011	-
CN	-	-	0.17	0.11	0.034

Table S6: Model selection results for *Sim2* in CA. Bold-numbered models (or their symmetric analogs) are consistent with the true data-generating model. Bold values indicate the best value in each column. Here, model 20 is chosen using the validation set, which is consistent with the data-generating model. It also has the best negative log-likelihood on the test set, though model 4 does slightly better on MSE of the class probabilities. All but one of the *FP* models outperform the *OCC* model, and they all outperform the *CN* model on ψ.

Model	Validation NLL	Test NLL	Test MSE on ψ	Test MSE on η	Test MSE on ρ
1	2113.0	2435.6	0.061	0.0011	0.0090
2	2113.7	2454.4	0.065	0.00030	**0.00024**
3	2061.4	2360.7	0.045	0.021	0.0094
4	2115.9	2438.1	0.061	0.0011	0.0089
5	2129.8	2444.6	0.062	0.0010	0.0090
6	2095.9	2469.3	0.062	0.027	0.00092
7	2043.8	**2317.4**	**0.040**	0.021	0.00046
8	2116.7	2456.3	0.066	0.00030	0.00028
9	2122.7	2456.5	0.066	0.00025	0.00056
10	2111.4	2507.3	0.067	0.028	0.0099
11	2107.7	2475.1	0.062	0.027	0.0098
12	2058.4	2374.7	0.047	0.021	0.0095
13	2124.6	2456.3	0.066	**0.00025**	0.00055
14	2129.7	2443.8	0.062	0.0010	0.0090
15	**2041.7**	2327.9	0.042	0.021	0.00083
16	2093.2	2440.4	0.058	0.027	0.00056
17	2143.9	2484.8	0.070	0.00036	0.0010
18	2144.6	2462.6	0.067	0.00048	0.0011
19	2121.8	2448.1	0.065	0.00051	0.00062
20	2143.8	2467.8	0.068	0.00051	0.0010
21	2144.2	2467.1	0.068	0.00051	0.0010
OCC	-	-	0.050	0.0014	-
CN	-	-	0.055	0.057	0.0095

Table S7: Model selection results for *Sim1* in NY. Bold-numbered models (or their symmetric analogs) are consistent with the true data-generating model. Bold values indicate the best value in each column. Here, model 15 is chosen using the validation set even though it is not consistent with the data-generating model. On the test set, model 7 performs best on the class model, even though it is not consistent with the data-generating model. Four of the *FP* models outperform the *OCC* and *CN* models on ψ.

Model	Validation NLL	Test NLL	Test MSE on ψ	Test MSE on η	Test MSE on ρ
1	2223.6	2132.6	0.037	0.0025	0.045
2	2072.1	2037.8	0.029	0.00017	0.023
3	2448.2	2367.0	0.11	0.042	0.046
4	2227.9	2141.3	0.039	0.0025	0.045
5	2223.6	2128.2	0.036	0.0025	0.044
6	2378.9	2310.5	0.076	0.053	0.025
7	2342.0	2265.4	0.080	0.039	0.023
8	2072.0	2042.6	0.030	0.00016	0.023
9	2076.0	2029.5	0.026	0.00012	0.022
10	2496.4	2424.7	0.036	0.085	0.0017
11	2482.7	2425.4	0.036	0.085	0.0017
12	2462.7	2378.5	0.11	0.042	0.046
13	2072.1	2033.9	0.028	**0.00011**	0.022
14	2225.5	2136.9	0.039	0.0024	0.045
15	2358.1	2272.1	0.080	0.040	0.023
16	2370.7	2305.0	0.076	0.051	0.025
17	2071.5	1957.9	0.0093	0.00013	0.00025
18	2071.8	**1957.7**	0.0091	0.00014	0.00024
19	2075.3	2037.4	0.029	0.00021	0.022
20	**2067.2**	1959.1	0.094	0.00016	**0.00015**
21	2071.4	1957.8	**0.0091**	0.00013	0.00025
OCC	-	-	0.034	0.016	-
CN	-	-	0.085	0.11	0.049

Table S8: Model selection results for *Sim2* in NY. Bold-numbered models (or their symmetric analogs) are consistent with the true data-generating model. Bold values indicate the best value in each column. Here, model 20 is chosen using the validation set, which is consistent with the data-generating model. Model 18 has the best negative log-likelihood on the test set, and model 21 performs best in terms of MSE on ψ for the test set, but all four consistent models have very similar performance. Eight of the *FP* models outperform the *OCC* model, and 19 of them outperform the *CN* model on ψ.