A Additional Information on Fair-OBNC Configurations

In this section of the appendix, we give additional details on possible configurations of the Fair-OBNC method. In our propsoed method, we provide the possible tunable parameters of the original OBNC method plus some additional configurable settings. These are:

- The ceiling for the number of instances to change classification, M. This can also be defined as a percentage of the training instances.
- The target disparity D. This represents the maximum ratio difference between the label prevalence of the most distant group to the average prevalence of the dataset.
- The set of features of the dataset to be ignored by the ensemble F. By default, $F = \{s\}$.
- The margin flip threshold t. This determines the minimum value of the margin function $N(x_i)$ for an instance to be considered as a candidate for flipping. Setting $t = 0$ results in only instances incorrectly classified by the algorithm being flipped $\phi(x_i) \neq y_i$, as in the original OBNC algorithm.
- The method for margin calculation $margin(x)$. This method can either use the vote (*i.e.*, the predicted label), or the score $\phi_n(x_i) \simeq P(y_i = 1|x_i)$ of each classifier n of the ensemble.
- The number of iterations of the bagging estimator, k.
- The bagging factor for the ensemble algorithm $\lambda = \frac{n'}{n}$, where n' is the size of the bootstrapped dataset and n is the size of the original dataset. This improves training times of the bagging estimator for large datasets.
- The classification algorithm to be used in the bagging estimator, such as Classification and Regression Trees (CART).
- The hyperparameters of the classification algorithm within the bagging estimator, such as the splitting criterion.

B Additional information on the experimental setup

B.1 Hyperparameter Configurations

For each of the considered pre-processing baselines, we run 50 rounds of hyperparameter optimization, using random search. This randomly samples a combination of the hyperparameters presented in the Tables 3 through 9, depending to the considered method. The hyperparameters for the LightGBM model being trained on the modified data are also randomly sampled from the configurations in Table 2.

Table 2. Model hyperparameter grid for LightGBM

Hyperparameter	Distribution	Values
boosting_type	-	"dart", "gbdt"
enable_bundle	-	*False*
n_estimators	Uniform	{10, 100}
min_child_samples	Log-uniform	{1, 500}
learning_rate	Uniform	[0.001, 0.1]

Table 3. Model hyperparameter grid for OBNC

Hyperparameter	Distribution	Values
max_flip_rate	Uniform	[0.0, 0.5]
bagging_max_samples	Uniform	[0.0, 1.0]
bagging_base_estimator	-	"LigthGBM"
bagging_n_estimators	Log-uniform	5, 30
boosting_type	-	"dart", "gbdt"

Table 4. Model hyperparameter grid for Fair-OBNC

Hyperparameter	Distribution	Values
max_flip_rate	Uniform	[0.0, 0.5]
bagging_max_samples	Uniform	[0.0, 1.0]
bagging_base_estimator	-	"LightGBM"
bagging_n_estimators	Log-uniform	5, 30
margin_type	-	"voting", "scores"
unawareness_features	-	*True*, *False*
boosting_type	-	"dart", "gbdt"

Table 5. Model hyperparameter grid for Massaging

Hyperparameter	Distribution	Values
classifier	-	"GaussianNB", "LightGBM"
boosting_type	-	"dart", "gbdt"

Table 6. Model hyperparameter grid for Prevalence Sampling

Hyperparameter	Distribution	Values
alpha	Uniform	[0.1, 1.0]
strategy	-	"undersample", "oversample"

Table 7. Model hyperparameter grid for Data Repairer

Hyperparameter	Distribution	Values
repair_level	Uniform	[0.1, 1.0]

Table 8. Model hyperparameter grid for Correlation Suppression

Hyperparameter	Distribution	Values
correlation_threshold	Uniform	[0.1, 0.9]

Table 9. Model hyperparameter grid for Feature Importance Suppression

Hyperparameter	Distribution	Values
auc_threshold	Uniform	[0.1, 0.9]
feature_importance_threshold	Uniform	[0.05, 0.5]
n_estimators	Uniform	10, 100

B.2 Dataset

In this section, we provide some insights about the distribution of the used dataset and how the performed noise injection alters this distribution. The tuples in the following tables indicate the label (Positive $+$, Negative $-$, or both $\cdot$) and group (older than 50 A, younger than 50 B, or both $\cdot$) of the segment analyzed. In Table 10 we present the distribution of the Variant II of the BankAccountFraud dataset,

followed by how it changes when we make the dataset IID, which is shown in Table 11.

Table 10. Original Bank Account Fraud (Variant II) dataset proportions.

Split	# instances	$(+, \cdot)$	$(\cdot, A)$	$(+, A)$	$(+, B)$
train	794990	1.0%	50.6%	1.7%	0.3%
validation	108168	1.3%	50.6%	2.2%	0.4%
test	96842	1.5%	50.5%	2.4%	0.5%

Table 11. IID version of Bank Account Fraud (Variant II) dataset.

Split	$(+, \cdot)$	$(\cdot, A)$	$(+, A)$	$(+, B)$
train	1.1%	50.5%	1.1%	1.1%
validation	1.1%	50.8%	1.1%	1.1%
test	1.1%	50.7%	1.1%	1.2%

Despite only presenting the case where noise is injected at increasing rates in the sensitive group A (instances with age above 50), we conduct experiments with multiple noise injection processes to better understand how the methods fare in different noise scenarios. The three considered noise types are as follows:

- *Type I*: Noise is injected at the same rate in instances of both sensitive groups simultaneously;
- *Type II*: The noise type considered in the main manuscript, where noise is injected solely on the instances that belong to the sensitive group A;
- *Type III*: A fixed rate of 5% of noise is applied to the sensitive group B while noise in the group A is progressively increased.

A visual representation of how the noise increases in each sensitive group for each noise type is presented in Fig. 4.

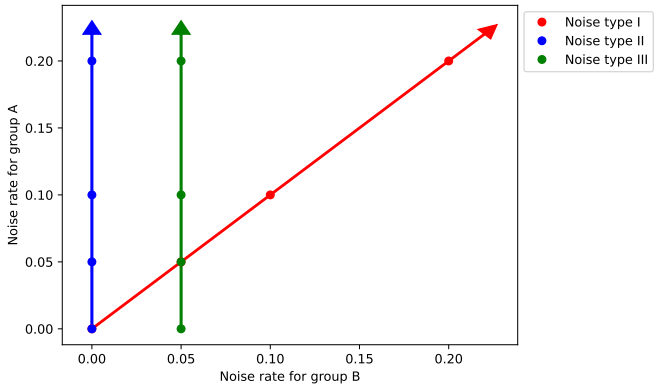

Figure 4. Visual representation of the considered types of noise.

Table 12 shows the distribution of every variation of the training set after injecting noise using all the previously described combinations.

B.3 Computational Setup

All experiments were conducted in 40 vCPU cores of Intel Xeon Gold 5120 @2.2GHz with 126GB of RAM. A total of 12.000 models were trained and evaluated. The average computing times are presented in Table 13.

Table 12. Noisy versions of IID Bank Account Fraud (Variant II) train set.

Noisy label	Noise rate ($s = B, s = A$)	$(+, \cdot)$	$(+, A)$	$(+, B)$
Label 0	(0%,0%)	1.1%	1.1%	1.1%
	(5%,5%)	6.0%	6.0%	6.0%
	(10%,10%)	11.0%	11.0%	11.0%
	(20%,20%)	20.9%	20.9%	20.9%
	(0%,5%)	3.4%	6.0%	1.1%
	(0%,10%)	6.1%	11.0%	1.1%
	(0%,20%)	11.1%	20.9%	1.1%
	(5%,0%)	3.5%	1.1%	6.1%
	(5%,10%)	8.5%	11.0%	6.1%
	(5%,20%)	13.5%	21.0%	6.1%
Label 1	(0%,0%)	1.1%	1.1%	1.1%
	(5%,5%)	1.0%	1.0%	1.1%
	(10%,10%)	1.0%	1.0%	1.0%
	(20%,20%)	0.9%	0.9%	0.9%
	(0%,5%)	1.1%	1.0%	1.1%
	(0%,10%)	1.0%	1.0%	1.1%
	(0%,20%)	1.0%	0.9%	1.1%
	(5%,0%)	1.1%	1.1%	1.1%
	(5%,10%)	1.0%	1.0%	1.1%
	(5%,20%)	1.0%	0.9%	1.1%
Both labels	(0%,0%)	1.1%	1.1%	1.1%
	(5%,5%)	6.0%	6.0%	6.0%
	(10%,10%)	10.9%	10.9%	10.9%
	(20%,20%)	20.7%	20.7%	20.7%
	(0%,5%)	3.6%	6.0%	1.1%
	(0%,10%)	6.0%	10.9%	1.1%
	(0%,20%)	11.0%	20.7%	1.1%
	(5%,0%)	3.5%	1.1%	6.0%
	(5%,10%)	8.5%	10.9%	6.0%
	(5%,20%)	13.4%	20.7%	6.0%

Table 13. Average computing times in seconds for each method used in the experiment.

Method	Time (s)
No preprocessing	84.7
OBNC	1321.8
Fair-OBNC	1241.8
Massaging	148.4
Prevalence Sampling	71.0
Data Repairer	95.4
Correlation Suppression	78.5
Feature Importance Suppression	681.8

C Additional results and discussion

In this section we present and elaborate on the results obtained from the full range of conducted experiments.

C.1 Evaluation of label reconstruction

Regarding the label noise correction methods, we present how these performed in terms of effectively detecting noisy labels and flipping them to their original value. These results are presented in Table 14 for noise of *Type I* and in Table 15 for noise of *Type III*.

C.2 *Evaluation of performance*

The additional results obtained from performing the experiments using all the noise types are presented in this section. The values in the following plots are obtained by averaging the performance (according to the considered metric) of each method over the 50 trials of the experiment, as described above.

C.2.1 *Type I*

For the negative label, the TPR and Demographic Parity observed over the considered noise rates are presented in Figs. 5 and 6, respectively.

Considering the case of noise injection in the positive label samples, the obtained results are presented in Figs. 7 and 8.

For both labels, these are shown in Figs. 9 and 10.

C.2.2 *Type II*

In this section, we show the results obtained when injecting noise at increasing rates in the positive label and on both labels, since the case where noise is injected in the negative label is presented in the main manuscript.

For the positive label, the TPR and Demographic Parity observed over the considered noise rates are presented in Figs. 11 and 12, respectively.

Considering the case where there is noise in both labels, the TPR results are presented in Fig. 13, and the Demographic Parity in Fig. 14.

C.2.3 *Type III*

Regarding noise injection on negative label instances, the TPR and Demographic Parity observed over the considered noise rates are presented in Figs. 15 and 16, respectively.

For the positive label, the TPR and Demographic Parity observed over the considered noise rates are presented in Figs. 17 and 18, respectively.

Finally, when considering both labels in noise injection, the obtained results are shown in Figs. 19 and 20.

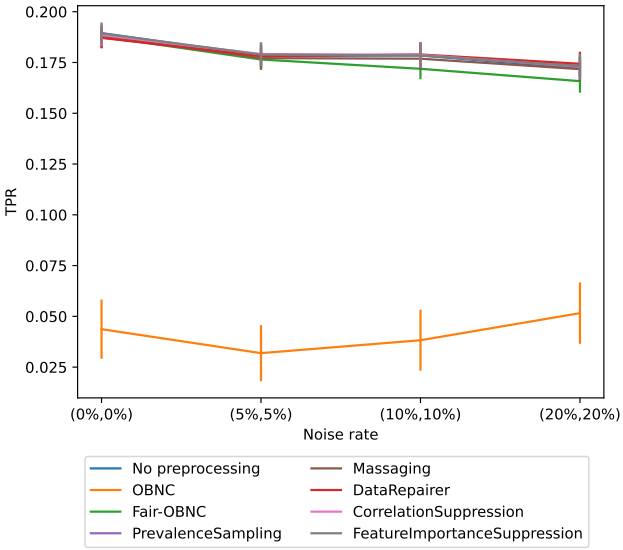

Figure 5. Average True Positive Rate achieved for each pre-processing method, for increasing noise rates of *Type I* noise injected in the negative label samples.

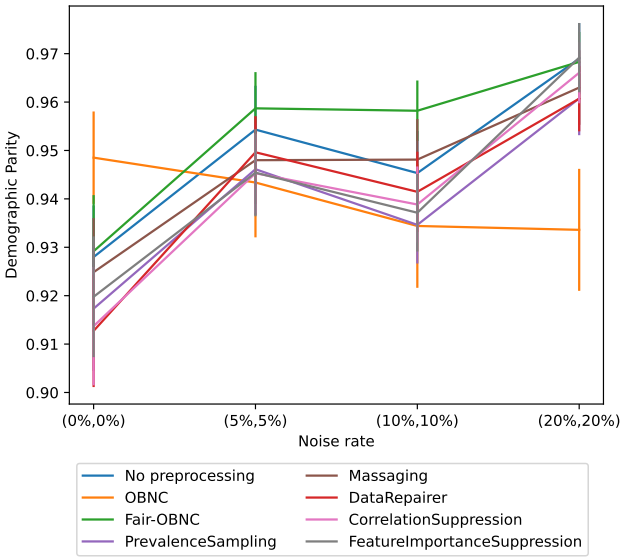

Figure 6. Average Demographic Parity achieved for each pre-processing method, for increasing noise rates of *Type I* noise injected in the negative label samples.

Table 14. Label noise correction performance across the noise rates for *Type I* noise.

		Noise rate: (5%, 5%)			Noise rate: (10%, 10%)			Noise rate: (20%, 20%)		
		OBNC	**Fair-OBNC**	**Massaging**	**OBNC**	**Fair-OBNC**	**Massaging**	**OBNC**	**Fair-OBNC**	**Massaging**
Label 0	Reconstruction Score	0.7728	0.9491	0.9505	0.8120	0.8984	0.9011	0.8615	0.7972	0.8022
	FPR	0.2193	0.0515	0.0500	0.1800	0.1027	0.1000	0.1310	0.2051	0.2000
	FNR	0.9380	0.0000	0.0000	0.9123	0.0000	0.0000	0.8160	0.0000	0.0000
	FPR (group B)	0.2189	0.0514	0.0499	0.1797	0.1026	0.0999	0.1307	0.2050	0.1999
	FNR (group B)	0.9379	0.0000	0.0000	0.9122	0.0000	0.0000	0.8156	0.0000	0.0000
	FPR (group A)	0.2198	0.0515	0.0501	0.1803	0.1027	0.1000	0.1312	0.2051	0.2000
	FNR (group A)	0.9380	0.0000	0.0000	0.9124	0.0000	0.0000	0.8164	0.0000	0.0000
Label 1	Reconstruction Score	0.7290	0.9992	0.9994	0.7290	0.9987	0.9988	0.7286	0.9976	0.9977
	FPR	0.2632	0.0003	0.0000	0.2635	0.0003	0.0000	0.2642	0.0002	0.0000
	FNR	0.9706	0.0489	0.0545	0.9504	0.0976	0.1040	0.9186	0.1963	0.2030
	FPR (group B)	0.2628	0.0003	0.0000	0.2631	0.0002	0.0000	0.2638	0.0002	0.0000
	FNR (group B)	0.9717	0.0493	0.0589	0.9513	0.0977	0.1080	0.9163	0.1961	0.2070
	FPR (group A)	0.2636	0.0004	0.0001	0.2639	0.0003	0.0001	0.2646	0.0003	0.0001
	FNR (group A)	0.9696	0.0486	0.0501	0.9496	0.0975	0.0997	0.9209	0.1965	0.1997
Both Labels	Reconstruction Score	0.7725	0.9486	0.9500	0.8115	0.8976	0.9000	0.8606	0.7955	0.8000
	FPR	0.2197	0.0514	0.0500	0.1806	0.1030	0.1000	0.1317	0.2050	0.2000
	FNR	0.9278	0.0470	0.0500	0.8978	0.0916	0.0998	0.8290	0.1793	0.2000
	FPR (group B)	0.2193	0.0514	0.0499	0.1803	0.1030	0.0999	0.1317	0.2050	0.1999
	FNR (group B)	0.9277	0.0475	0.0503	0.8984	0.0912	0.1000	0.8273	0.1796	0.2000
	FPR (group A)	0.2201	0.0515	0.0501	0.1809	0.1030	0.1000	0.1318	0.2050	0.2000
	FNR (group A)	0.9279	0.0465	0.0497	0.8973	0.0919	0.0995	0.8306	0.1791	0.1991

Table 15. Label noise correction performance across the noise rates for *Type III* noise.

		Noise rate: (5%, 5%)			Noise rate: (5%, 10%)			Noise rate: (5%, 20%)		
		OBNC	**Fair-OBNC**	**Massaging**	**OBNC**	**Fair-OBNC**	**Massaging**	**OBNC**	**Fair-OBNC**	**Massaging**
Label 0	Reconstruction Score	0.7728	0.9491	0.9505	0.7937	0.9299	0.9253	0.8286	0.8930	0.8751
	FPR	0.2193	0.0515	0.0500	0.1983	0.0704	0.0754	0.1634	0.1072	0.1260
	FNR	0.9380	0.0000	0.0000	0.9278	0.0473	0.0114	0.8900	0.0848	0.0212
	FPR (group B)	0.2189	0.0514	0.0499	0.2012	0.0598	0.0752	0.1677	0.0780	0.1257
	FNR (group B)	0.9379	0.0000	0.0000	0.9275	0.0000	0.0000	0.8875	0.0000	0.0000
	FPR (group A)	0.2198	0.0515	0.0501	0.1955	0.0807	0.0756	0.1591	0.1358	0.1260
	FNR (group A)	0.9380	0.0000	0.0000	0.9281	0.0944	0.0228	0.8925	0.1692	0.0423
Label 1	Reconstruction Score	0.7290	0.9992	0.9994	0.7289	0.9989	0.9988	0.7288	0.9983	0.9978
	FPR	0.2632	0.0003	0.0000	0.2634	0.0003	0.0002	0.2637	0.0001	0.0004
	FNR	0.9706	0.0489	0.0545	0.9628	0.0732	0.0913	0.9454	0.1428	0.1640
	FPR (group B)	0.2628	0.0003	0.0000	0.2631	0.0001	0.0000	0.2637	0.0000	0.0000
	FNR (group B)	0.9717	0.0493	0.0589	0.9709	0.0492	0.0835	0.9718	0.0899	0.1330
	FPR (group A)	0.2636	0.0004	0.0001	0.2637	0.0005	0.0004	0.2637	0.0002	0.0009
	FNR (group A)	0.9696	0.0486	0.0501	0.9546	0.0971	0.0991	0.9191	0.1955	0.1948
Both Labels	Reconstruction Score	0.7725	0.9486	0.9500	0.7933	0.9292	0.9246	0.8281	0.8919	0.8740
	FPR	0.2197	0.0514	0.0500	0.1988	0.0704	0.0753	0.1641	0.1070	0.1260
	FNR	0.9278	0.0470	0.0500	0.9145	0.1129	0.0795	0.8808	0.1896	0.1320
	FPR (group B)	0.2193	0.0514	0.0499	0.2019	0.0597	0.0748	0.1686	0.0776	0.1247
	FNR (group B)	0.9277	0.0475	0.0503	0.9185	0.0426	0.0387	0.8853	0.0394	0.0273
	FPR (group A)	0.2201	0.0515	0.0501	0.1958	0.0808	0.0759	0.1596	0.1360	0.1270
	FNR (group A)	0.9279	0.0465	0.0497	0.9104	0.1829	0.1200	0.8764	0.3393	0.2358

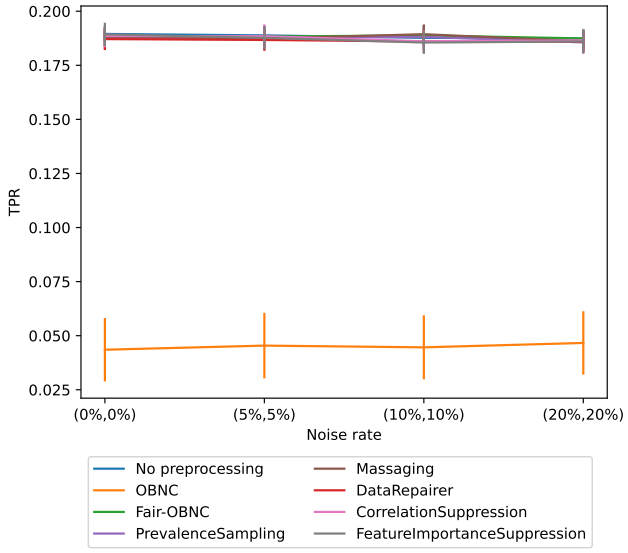

Figure 7. Average True Positive Rate achieved for each pre-processing method, for increasing noise rates of *Type I* noise injected in the positive label samples.

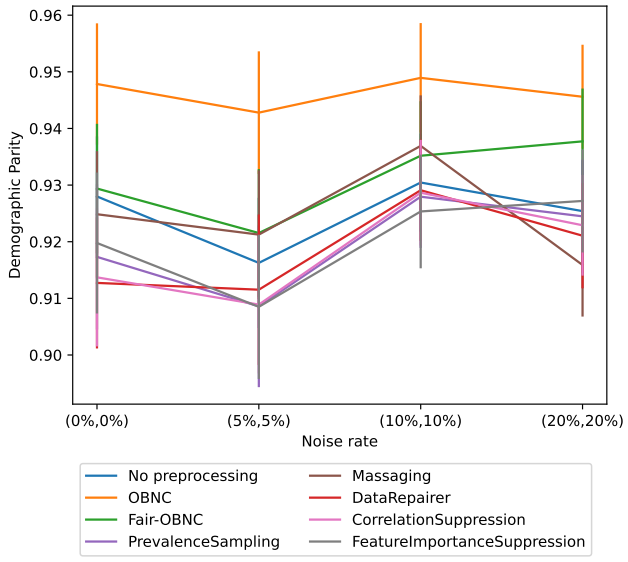

Figure 8. Average Demographic Parity achieved for each pre-processing method, for increasing noise rates of *Type I* noise injected in the positive label samples.

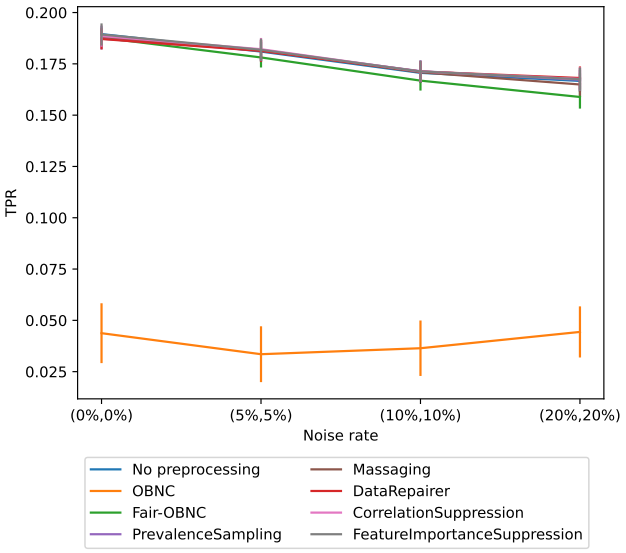

Figure 9. Average True Positive Rate achieved for each pre-processing method, for increasing noise rates of *Type I* noise.

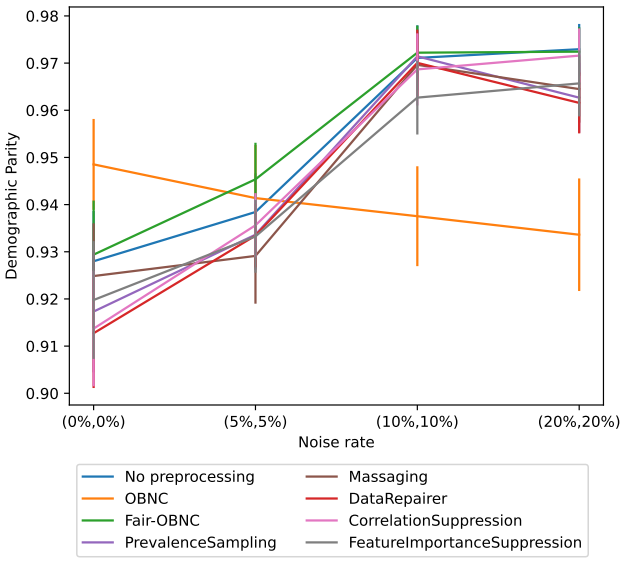

Figure 10. Average Demographic Parity achieved for each pre-processing method, for increasing noise rates of *Type I* noise.

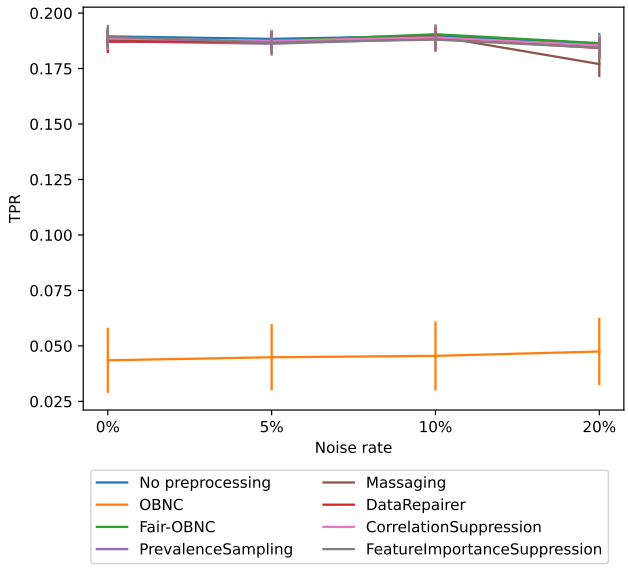

Figure 11. Average True Positive Rate achieved for each pre-processing method, for increasing noise rates of *Type II* noise injected in the positive label samples.

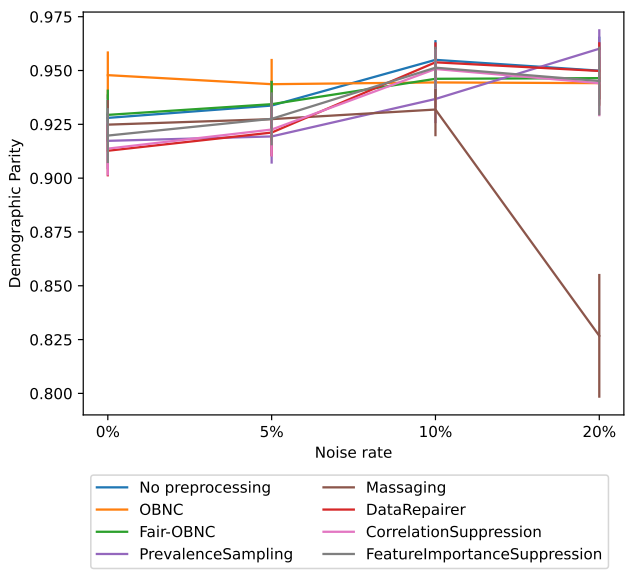

Figure 12. Average Demographic Parity achieved for each pre-processing method, for increasing noise rates of *Type II* noise injected in the positive label samples.

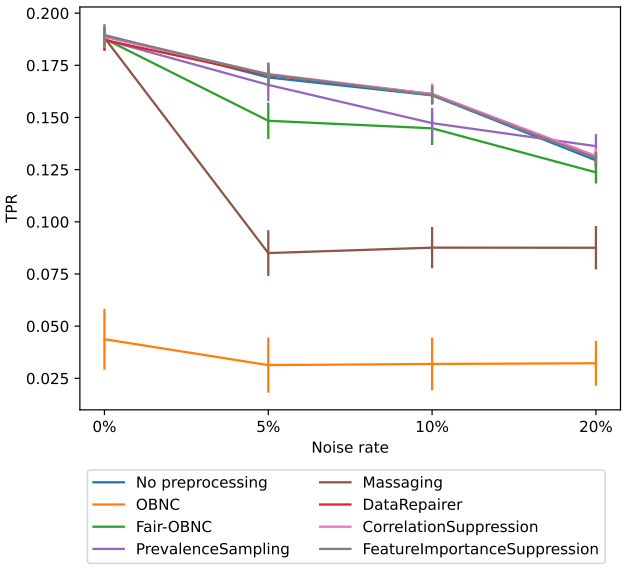

Figure 13. Average True Positive Rate achieved for each pre-processing method, for increasing noise rates of *Type II* noise.

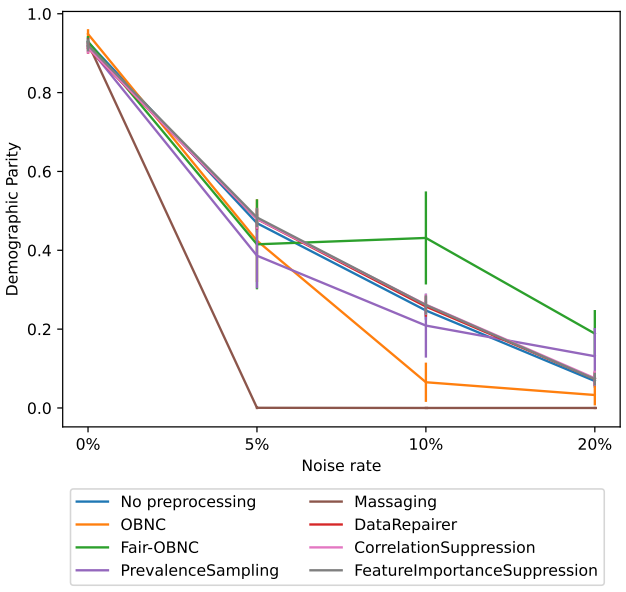

Figure 14. Average Demographic Parity achieved for each pre-processing method, for increasing noise rates of *Type II* noise.

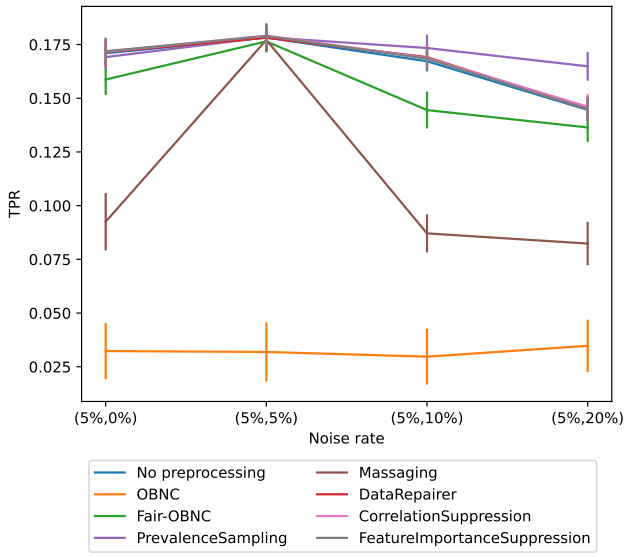

Figure 15. Average True Positive Rate achieved for each pre-processing method, for increasing noise rates of *Type III* noise injected in the negative label samples.

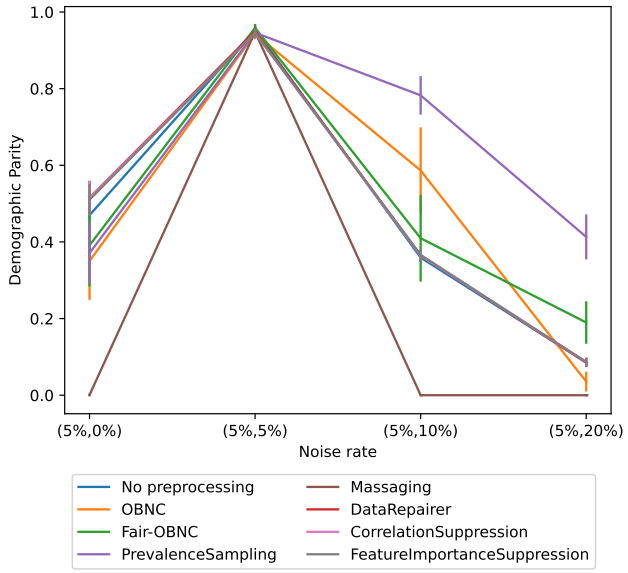

Figure 16. Average Demographic Parity achieved for each pre-processing method, for increasing noise rates of *Type III* noise injected in the negative label samples.

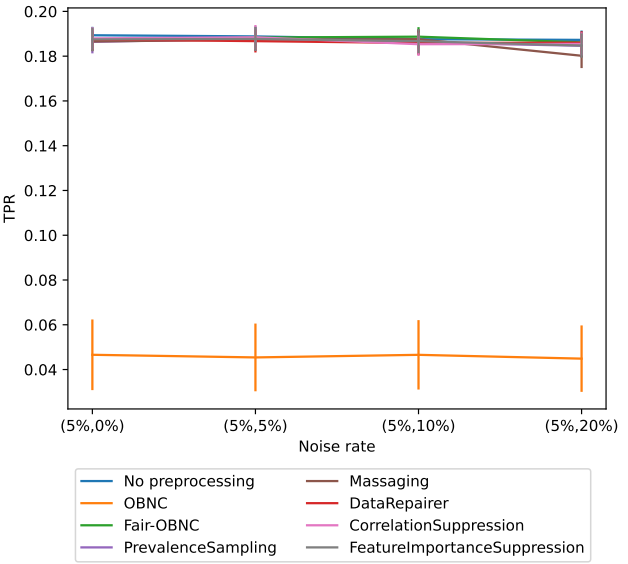

Figure 17. Average True Positive Rate achieved for each pre-processing method, for increasing noise rates of *Type III* noise injected in the positive label samples.

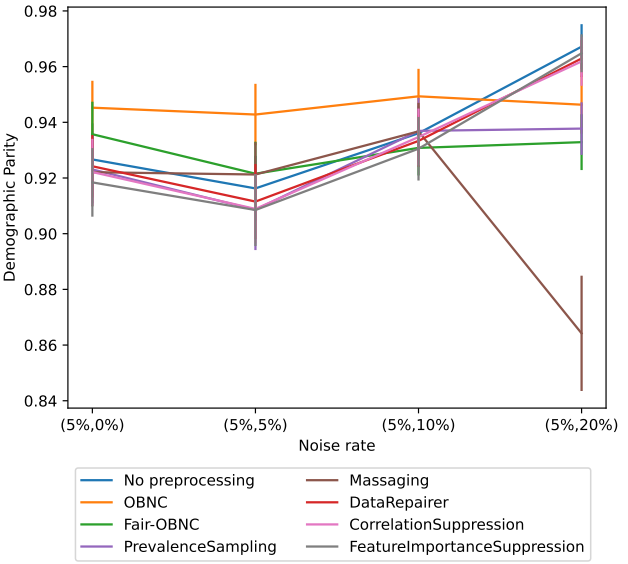

Figure 18. Average Demographic Parity achieved for each pre-processing method, for increasing noise rates of *Type III* noise injected in the positive label samples.

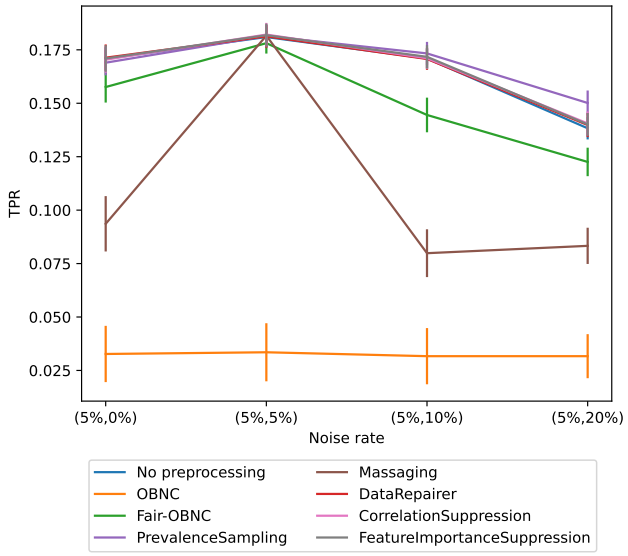

Figure 19. Average True Positive Rate achieved for each pre-processing method, for increasing noise rates of *Type III* noise.

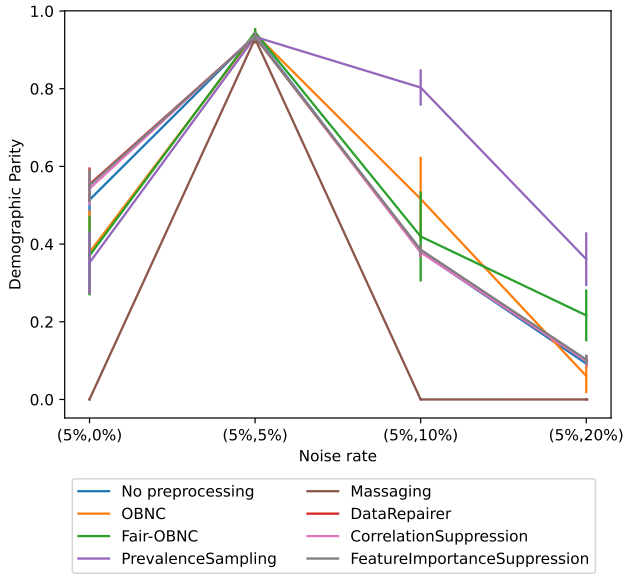

Figure 20. Average Demographic Parity achieved for each pre-processing method, for increasing noise rates of *Type III* noise.