Appendix

This document presents some supplementary material to the manuscript submitted to Artificial Intelligence Review, entitled “On the joint-effect of Class Imbalance and Overlap: A Critical Review”. **Authors:** *Miriam Seoane Santos, Pedro Henriques Abreu, Nathalie Japkowicz, Alberto Fernández, Carlos Soares, Szymon Wilk, and João Santos.*

A Lessons learned (supporting information)

Table A.1: Characterisation of the behaviour of classifiers from related work. In this table are included the typical and atypical domains from García et al. [2,4,5,3] and the domains by Prati et al. [8] and Denil and Trappenberg [1].

Typical Domains: Squares, IR = 4:1			Atypical Domains: Squares, IR = 4:1		
Classifier	**Sensitivity**	**Specificity**	**Classifier**	**Sensitivity**	**Specificity**
KNN [2,4] [5,3]	Sensitivity of 50%, 30% and 20% for higher percentages of class overlap (60%, 80% and 100% respectively) for 1NN. Faster deterioration was reported for higher values of k ($k = 3$, $k = 9$) [3].	Specificity decreases (100% to 80%) as overlap increases (from 0% to 100%) for 1NN. Higher values of k seem to benefit the majority class: specificity around 100% to 90% for 0% to 100% overlap for $k = 3$ and stable at 100% for $k = 9$ [3].	**KNN** [4,5,3]	Sensitivity increases as the minority class gets denser (40% to 80%). Increasing the value ok k benefits the minority class (range of 40% to 90% for $k = 3$ and 40% to 100% for $k = 9$) [3].	Specificity stable around 80%-95% as the minority gets denser. Specificity is always superior to Sensitivity. Increasing the value of k does not seem to impact the results [3].
MLP [4,5,3]	Sensitivity around 40%, 20% and 0% for higher percentages of class overlap (60%, 80% and 100% respectively)	Specificity remains stable (near 100%) as overlap increases.	**MLP** [4,5,3]	Sensitivity increases as the minority class gets denser (40% to 100%). Sensitivity and specificity start apart for the balanced configuration (40% and 80% respectively) and go hand-in-hand as the minority class becomes denser (80% to 100%).	Specificity stable around 80%-95% as the minority gets denser. Shows an inflection curve where the specificity decreases for the first configuration where classes interchange roles (from the balanced configuration [75-100] to the [80-100] configuration), before starting to increase gradually.
C4.5 [4,5,3]	Sensitivity around 40%, 20% and 0% for higher percentages of class overlap (60%, 80% and 100% respectively)	Specificity remains stable (near 100%) as overlap increases.	**C4.5** [4,5,3]	Sensitivity increases as the minority class gets denser (40% to 100%). Sensitivity and specificity are considerably different for the balanced configuration (40% / 80%), yet sensitivity rapidly increases to 100% in the following configurations, while specificity increases gradually.	Specificity stable around 80%-95% as the minority gets denser. Shows an inflection curve where the specificity decreases for the first configuration where classes interchange roles (from the balanced configuration [75-100] to the [80-100] configuration), before starting to increase gradually.
RBF [4,5,3]	Sensitivity around 40%, 20% and 0% for higher percentages of class overlap (60%, 80% and 100% respectively)	Specificity remains stable (near 100%) as overlap increases. Nevertheless, a slight decrease is noticeable for intermediate levels of overlap (around 2%).	**RBF** [4,5,3]	Sensitivity increases as the minority class gets denser (40% to 100%) but only surpasses specificity for the final configuration, [95-100], and increases slowly.	Specificity stable around 80%-95% as the minority gets denser.
SVM [5]	Sensitivity of 50% for 40% overlap and 0% for higher overlap levels (from 60% to 100%).	Specificity remains stable (near 100%) as overlap increases.	**SVM** [5]	Sensitivity increases as the minority class gets denser, although very slowly: 0% for the [75-100] (balanced) and [80-100] configurations, and 20% for [85-100]. For the final two configurations, sensitivity rises to 90% and 100%.	Specificity decreases as the minority class gets denser, although slightly (100% to 90%).

To be continued on the next page. . .

Table A.1: Continued from previous page.

Classifier	Sensitivity	Specificity	Classifier	Sensitivity	Specificity
Typical Domains: Squares, IR 4:1			**Atypical Domains: Squares, IR = 4:1**		
NB [4,5,3]	Sensitivity around 40%, 20% and 0% for higher percentages of class overlap (60%, 80% and 100% respectively). A fast decrease is noted for class overlap over 60%: sensitivity below 20% was reported for 80% overlap citeGarcia2007b.	Specificity remains stable (near 100%) as overlap increases.	**NB** [4,5,3]	Sensitivity increases as the minority class gets denser (80% to 100%). For a balanced configuration, both classes present similar recognition rates (around 80%) and as the minority class gets denser, sensitivity assumes higher (although close) values than specificity.	Specificity stable around 80%-95% as the minority gets denser.
Atypical Domains: Concentric Circles, IR = 50:1			**Other Domains**		
KNN [3] **RBF** [3]	Sensitivity results are similar to standard atypical situations.		**C4.5** [8]	For 1 and 3 SD, C4.5 achieved an AUC of: 91% and 99.9% (IR = 4:1, 5D) 87% and 99.6% (IR = 9:1, 5D)	
C4.5 [3]	Sensitivity results are similar to standard atypical situations, although the performance for balanced configurations is lower in this domain (around 10%).	Specificity stable on 100%. For KNN, increasing the value of k does not seem to impact the results.	**SVM** [1]	SVM is capable of finding parsimonious models in the presence of class imbalance, whereas class overlap severely increases model complexity. When domains are both imbalanced and overlapped, SVM revealed a breaking point for $\alpha = 0.6$ (IR = 1.5) and $\mu = 0.78$.	
MLP [3]	Sensitivity of 0% for all configurations.				
NB [3]	Sensitivity of 100% for all configurations.				

Table A.2: Characterisation of the behaviour of classifiers from related work (*subclus* and *paw* domains).

Subclus Domains			Paw Domains		
Classifier	**Sensitivity**	**G-mean**	**Classifier**	**Sensitivity**	**G-mean**
MODLEM [7]	Sensitivity of 88%, 56%, 34% and 20% for 0%, 30%, 50% and 70% of borderline minority examples (IR = 7:1 and 5 subregions).	G-mean of 94%, 73%, 56% and 41% for 0%, 30%, 50% and 70% of borderline minority examples (IR = 7:1 and 5 subregions).	**MODLEM** [7]	Sensitivity of 83%, 61%, 45% and 29% for 0%, 30%, 50% and 70% of borderline minority examples (IR = 7:1 and 3 subregions).	G-mean of 90%, 76%, 66% and 51% for 0%, 30%, 50% and 70% of borderline minority examples (IR = 7:1 and 3 subregions).
C4.5 [7,9]	Sensitivity of 95%, 45%, 17% and 0% for 0%, 30%, 50% and 70% of borderline minority examples (IR = 7:1 and 5 subregions) [7].	G-mean of 97%, 65%, 35% and 0% for 0%, 30%, 50% and 70% of borderline minority examples (IR = 7:1 and 5 subregions) [7].	**C4.5** [7]	Sensitivity of 52%, 26%, 18% and 0.6% for 0%, 30%, 50% and 70% of borderline minority examples (C4.5, IR = 7:1 and 3 subregions) [7].	G-mean of 67%, 33%, 32% and 1.5% for 0%, 30%, 50% and 70% of borderline minority examples (C4.5, IR = 7:1 and 3 subregions) [7].
	Sensitivity results for 0%, 10% and 20% of borderline minority examples [9]: 96%, 91% and 85% (IR = 5:1 and 3 subregions) 94%, 90% and 75% (IR = 9:1 and 3 subregions) 96%, 87% and 76% (IR = 5:1 and 5 subregions) 90%, 81% and 66% (IR = 9:1 and 5 subregions)		**C4.5-P** [10] **C4.5-U** [10]	Sensitivity of 90% and 91% (C4.5-P) and 89% and 90% (C4.5-U) for 0% and 30% of borderline minority examples (IR = 7:1, 3 subregions, 3D) [10].	G-mean of 94% and 95% (C4.5-P) and 94% (C4.5-U) for 0% and 30% of borderline minority examples (IR = 7:1, 3 subregions, 3D) [10].
CART [6]	Sensitivity results for CART with 0% and 50% of borderline minority examples: 98% and 90% (IR = 4:1 and 5 subregions) 93% and 73% (IR = 10:1 and 5 subregions) 97% and 97% (IR = 4:1 and 5 subregions, 5D) 96% and 89% (IR = 10:1 and 5 subregions, 5D)		**PART-P** [1 **PART-U** [1	Sensitivity of 90% and 91% (PART-P) and 89% and 90% (PART-U) for 0% and 30% of borderline minority examples (IR = 7:1, 3 subregions, 3D).	G-mean of 92% and 93% (PART-P) and 94% and 93% (PART-U) for 0% and 30% of borderline minority examples (IR = 7:1, 3 subregions, 3D).
SVM [6]	For 0% and 50% of borderline minority examples SVM achieved a sensitivity of: Linear kernel: 48% and 40% (IR = 4:1 and 5 subregions) Linear kernel: 33% and 12% (IR = 10:1 and 5 subregions) RBF kernel: 90% and 85% (IR = 4:1 and 5 subregions) RBF kernel: 69% and 54% (IR =10:1 and 5 subregions) Linear kernel: 48% and 47% (IR = 4:1 and 5 subregions, 5D) Linear kernel: 41% and 35% (IR = 10:1 and 5 subregions, 5D) RBF kernel: 96% and 94% (IR = 4:1 and 5 subregions, 5D) RBF kernel: 84% and 75% (IR = 10:1 and 5 subregions, 5D)		**SVM** [10]	Sensitivity of 98% and 99% for 0% and 30% borderline minority examples (IR = 7:1, 3 subregions, 3D).	G-mean of 99% for 0% and 30% borderline minority examples (IR = 7:1, 3 subregions, 3D).
KNN [6]	For 0% and 50% of borderline minority examples KNN achieved a sensitivity of: 85% and 66% (IR = 4:1 and 5 subregions) 65% and 48% (IR = 10:1 and 5 subregions) 99% and 97% (IR = 4:1 and 5 subregions, 5D) 83% and 78% (IR = 10:1 and 5 subregions, 5D)		**KNN** [10]	Sensitivity of 95% for 0% and 30% borderline minority examples (IR = 7:1, 3 subregions, 3D). Increasing the value of k seems to improve sensitivity results.	G-mean of 97% and 96% for 0% and 30% borderline minority examples (IR = 7:1, 3 subregions, 3D). Increasing the value of k seems to improve G-mean results.
NB [6]	For 0% and 50% of borderline minority examples NB achieved a sensitivity of: 53% and 46% (IR = 4:1 and 5 subregions) 0% and 0% (IR = 10:1 and 5 subregions) 100% and 100% (IR = 4:1 and 5 subregions, 5D) 96% and 93% (IR = 10:1 and 5 subregions 5D)		**NB** [10]	Sensitivity of 87% and 88% for 0% and 30% borderline minority examples (IR = 7:1, 3 subregions, 3D).	G-mean of 92% for 0% and 30% borderline minority examples (IR = 7:1, 3 subregions, 3D).

To be continued on the next page...

Table A.2: Continued from previous page.

Subclus Domains			**Paw Domains**		
Classifier	**Sensitivity**	**G-mean**	**Classifier**	**Sensitivity**	**G-mean**
MLP [6]	For 0% and 50% of borderline minority examples MLP achieved a sensitivity of: 80% and 0% (IR = 4:1 and 5 subregions) 81% and 57% (IR = 10:1 and 5 subregions) 89% and 83% (IR = 4:1 and 5 subregions, 5D) 77% and 69% (IR = 10:1 and 5 subregions, 5D)		**RBF** [10]	Sensitivity of 95% and 94% for 0% and 30% borderline minority examples (IR = 7:1, 3 subregions, 3D).	G-mean of 97% and 96% for 0% and 30% borderline minority examples (IR = 7:1, 3 subregions, 3D).
FLD [6]	For 0% and 50% of borderline minority examples FLD achieved a sensitivity of: 0% and 0% (IR = 4:1 and 5 subregions) 0% and 0% (IR = 10:1 and 5 subregions) 0% and 0% (IR = 4:1 and 5 subregions, 5D) 0% and 0% (IR = 10:1 and 5 subregions, 5D)				

Table A.3: Characterisation of the behaviour of classifiers from related work (*clover*/*flower* domains).

Clover/Flower Domains			**Clover/Flower Domains**		
Classifier	**Sensitivity**	**G-mean**	**Classifier**	**Sensitivity**	**G-mean**
KNN [10,6]	Sensitivity of 98% for 0% and 30% borderline minority examples (1NN, IR = 7:1, 5 subregions, 3D). Increasing the value of k seems to provide higher sensitivity results [10]. Sensitivity results for 0% and 50% of borderline minority examples [6]: 91% and 79% (IR = 4:1 and 5 subregions) 66% and 49% (IR = 10:1 and 5 subregions) 100% and 100% (IR = 4:1 and 5 subregions, 5D) 100% and 99% (IR = 10:1 and 5 subregions, 5D)	G-mean of 98% for 0% and 30% borderline minority examples (1NN, IR = 7:1, 5 subregions, 3D). Increasing the value of k seems to improve G-mean results [10].	**C4.5** [7]	Sensitivity of 43%, 13%, 5% and 0.8% for 0%, 30%, 50% and 70% of borderline minority examples (C4.5, IR = 7:1 and 5 subregions) [7].	G-mean of 64%, 26%, 11% and 2% for 0%, 30%, 50% and 70% of borderline minority examples (C4.5, IR = 7:1 and 5 subregions) [7].
			C4.5-P [10] **C4.5-U** [10]	Sensitivity of 93% and 94% (C4.5-P) and 90% and 91% (C4.5-U) for 0% and 30% of borderline minority examples (IR = 7:1, 5 subregions, 3D [10].	G-mean of 96% (C4.5-P) and 94% and 95% (C4.5-U) for 0% and 30% of borderline minority examples (IR = 7:1, 5 subregions, 3D [10].
FLD [6]	For 0% and 50% of borderline minority examples FLD achieved a sensitivity of: 0% and 0% (IR = 4:1 and 5 subregions) 0% and 0% (IR = 10:1 and 5 subregions) 0% and 0% (IR = 4:1 and 5 subregions, 5D) 0% and 0% (IR = 10:1 and 5 subregions, 5D)		**MLP** [6]	For 0% and 50% of borderline minority examples MLP obtained a sensitivity of: 93% and 91% (IR = 4:1 and 5 subregions) 79% and 74% (IR = 10:1 and 5 subregions) 100% and 99% (IR = 4:1 and 5 subregions, 5D) 99% and 99% (IR = 10:1 and 5 subregions, 5D)	
CART [6]	Sensitivity results for 0% and 50% of borderline minority examples: 78% and 73% (IR = 4:1 and 5 subregions) 66% and 36% (IR = 10:1 and 5 subregions) 98% and 98% (IR = 4:1 and 5 subregions, 5D) 94% and 96% (IR = 10:1 and 5 subregions, 5D)		**RBF** [10]	Sensitivity of 93% and 98% for 0% and 30% borderline minority examples (IR = 7:1, 5 subregions, 3D).	G-mean of 96% and 99% for 0% and 30% borderline minority examples (IR = 7:1, 5 subregions, 3D).
PART-P [10 **PART-U** [10	Sensitivity of 92% (PART-P) and 90% (PART-U) for 0% and 30% borderline minority examples (IR = 7:1, 5 subregions, 3D).	G-mean of 95% (PART-P) and 94% (PART-U) for 0% and 30% borderline minority examples (IR = 7:1, 5 subregions, 3D).	**MODLEM** [7	Sensitivity of 57%, 43%, 28% and 21% for 0%, 30%, 50% and 70% of borderline minority examples (IR = 7:1 and 5 subregions).	G-mean of 74%, 64%, 51% and 42% for 0%, 30%, 50% and 70% of borderline minority examples (IR = 7:1 and 5 subregions).

To be continued on the next page. . .

Table A.3: Continued from previous page.

Clover/Flower Domains			Clover/Flower Domains		
Classifier	**Sensitivity**	**G-mean**	**Classifier**	**Sensitivity**	**G-mean**
NB [10,6]	Sensitivity of 99% for 0% and 30% borderline minority examples (IR = 7:1, 5 subregions, 3D) [10].	G-mean of 98% for 0% and 30% borderline minority examples (IR = 7:1, 5 subregions, 3D) [10].	**SVM** [10,6]	Sensitivity of 100% and 99% for 0% and 30% borderline minority examples (IR = 7:1, 5 subregions, 3D) [10].	G-mean of 100% and 99% for 0% and 30% borderline minority examples (IR = 7:1, 5 subregions, 3D) [10].
	Sensitivity results for 0% and 50% of borderline minority examples [6]: 23% and 18% (IR = 4:1 and 5 subregions) 0% and 0% (IR = 10:1 and 5 subregions) 100% and 100% (IR = 4:1 and 5 subregions, 5D) 100% and 100% (IR = 10:1 and 5 subregions, 5D)			Sensitivity results for 0% and 50% of borderline minority examples [6]: Linear kernel: 47% and 31% (IR = 4:1 and 5 subregions) Linear kernel: 46% and 40% (IR = 10:1 and 5 subregions) RBF kernel: 95% and 92% (IR = 4:1 and 5 subregions) RBF kernel: 88% and 66% (IR =10:1 and 5 subregions) Linear kernel: 36% and 21% (IR = 4:1 and 5 subregions, 5D) Linear kernel: 15% and 19% (IR = 10:1 and 5 subregions, 5D) RBF kernel: 100% and 99% (IR = 4:1 and 5 subregions, 5D) RBF kernel: 100% and 100% (IR =10:1 and 5 subregions, 5D)	

References

1. Denil, M., Trappenberg, T.: Overlap versus imbalance. In: Canadian Conference on Artificial Intelligence. pp. 220–231. Springer (2010)
2. García, V., Alejo, R., Sánchez, J., Sotoca, J., Mollineda, R.: Combined effects of class imbalance and class overlap on instance-based classification. In: International Conference on Intelligent Data Engineering and Automated Learning. pp. 371–378. Springer (2006)
3. García, V., Mollineda, R., Sánchez, J.: On the k-nn performance in a challenging scenario of imbalance and overlapping. Pattern Analysis and Applications 11(3-4), 269–280 (2008)
4. García, V., Mollineda, R., Sánchez, J., Alejo, R., Sotoca, J.: When overlapping unexpectedly alters the class imbalance effects. In: Iberian Conference on Pattern Recognition and Image Analysis. pp. 499–506. Springer (2007)
5. García, V., Sánchez, J., Mollineda, R.: An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In: Iberoamerican Congress on Pattern Recognition. pp. 397–406. Springer (2007)
6. Mercier, M., Santos, M., Abreu, P., Soares, C., Soares, J., Santos, J.: Analysing the footprint of classifiers in overlapped and imbalanced contexts. In: International Symposium on Intelligent Data Analysis. pp. 200–212. Springer (2018)
7. Napierała, K., Stefanowski, J., Wilk, S.: Learning from imbalanced data in presence of noisy and borderline examples. In: International Conference on Rough Sets and Current Trends in Computing. pp. 158–167. Springer (2010)
8. Prati, R., G., B., Monard, M.: Class imbalances versus class overlapping: an analysis of a learning system behavior. In: Mexican international conference on artificial intelligence. pp. 312–321. Springer (2004)
9. Stefanowski, J.: Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. In: Emerging paradigms in machine learning, pp. 277–306. Springer (2013)
10. Wojciechowski, S., Wilk, S.: Difficulty factors and preprocessing in imbalanced data sets: an experimental study on artificial data. Foundations of Computing and Decision Sciences 42(2), 149–176 (2017)