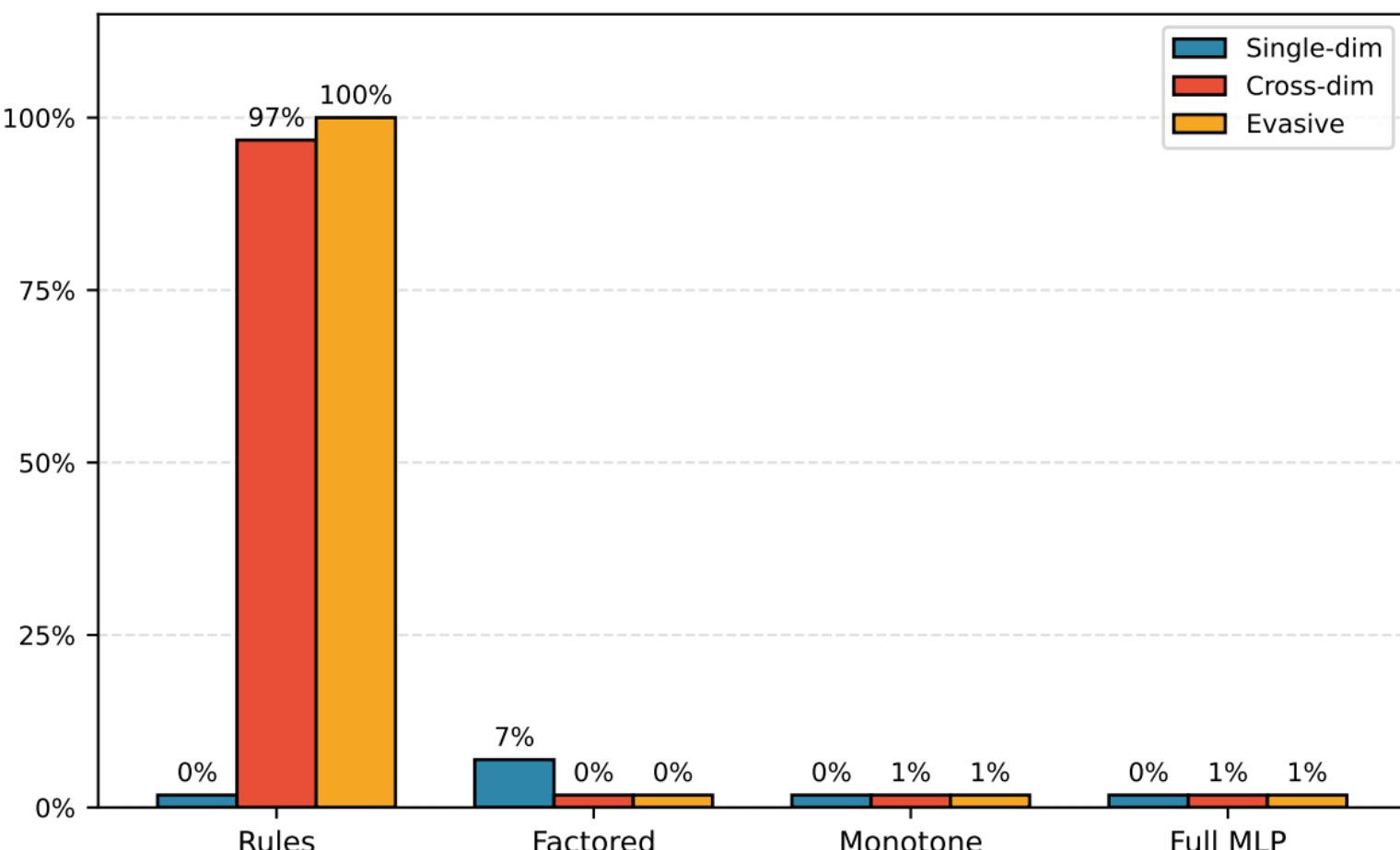
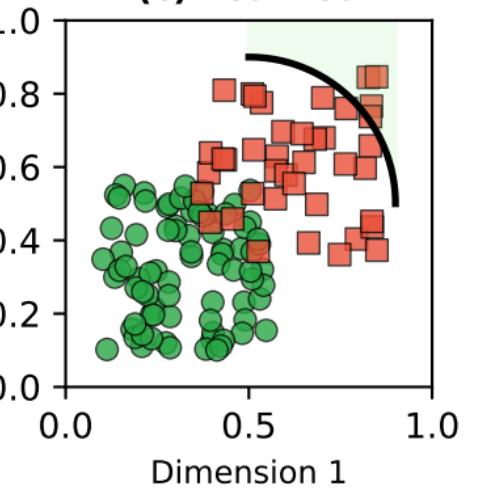
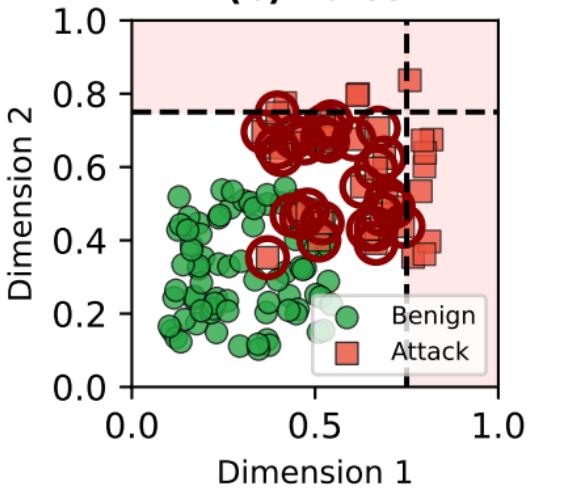


(a) False Accept Rate by Attack Type**(c) Learned****(b) Rules****(d) Summary**

Model	FAR	FRR	Acc	ECE
Rules	0.65	0.00	0.61	0.33
Factored	0.02	0.25	0.88	0.23
Monotone	0.01	0.01	0.99	0.00
Full MLP	0.01	0.01	0.99	0.01