

Attack	Classifier Model	No Attack	No Defense	Defense- GAN-Rec	MagNet	Adv. Tr. $\epsilon = 0.3$
FGSM $\epsilon = 0.3$	A	0.934	0.102	0.879	0.089	<u>0.797</u>
	B	0.747	0.102	0.629	<u>0.168</u>	0.136
	C	0.933	0.139	0.896	0.110	<u>0.804</u>
	D	0.892	0.082	0.875	0.099	<u>0.698</u>
RAND+FGSM $\epsilon = 0.3, \alpha = 0.05$	A	0.934	0.102	0.888	0.096	<u>0.447</u>
	B	0.747	0.131	0.661	<u>0.161</u>	0.119
	C	0.933	0.105	0.893	0.112	<u>0.699</u>
	D	0.892	0.091	0.862	0.104	<u>0.626</u>
CW ℓ_2 norm	A	0.934	0.076	0.896	0.060	<u>0.157</u>
	B	0.747	<u>0.172</u>	0.656	0.131	0.118
	C	0.933	0.063	0.896	0.084	<u>0.107</u>
	D	0.892	0.090	0.875	0.069	<u>0.149</u>