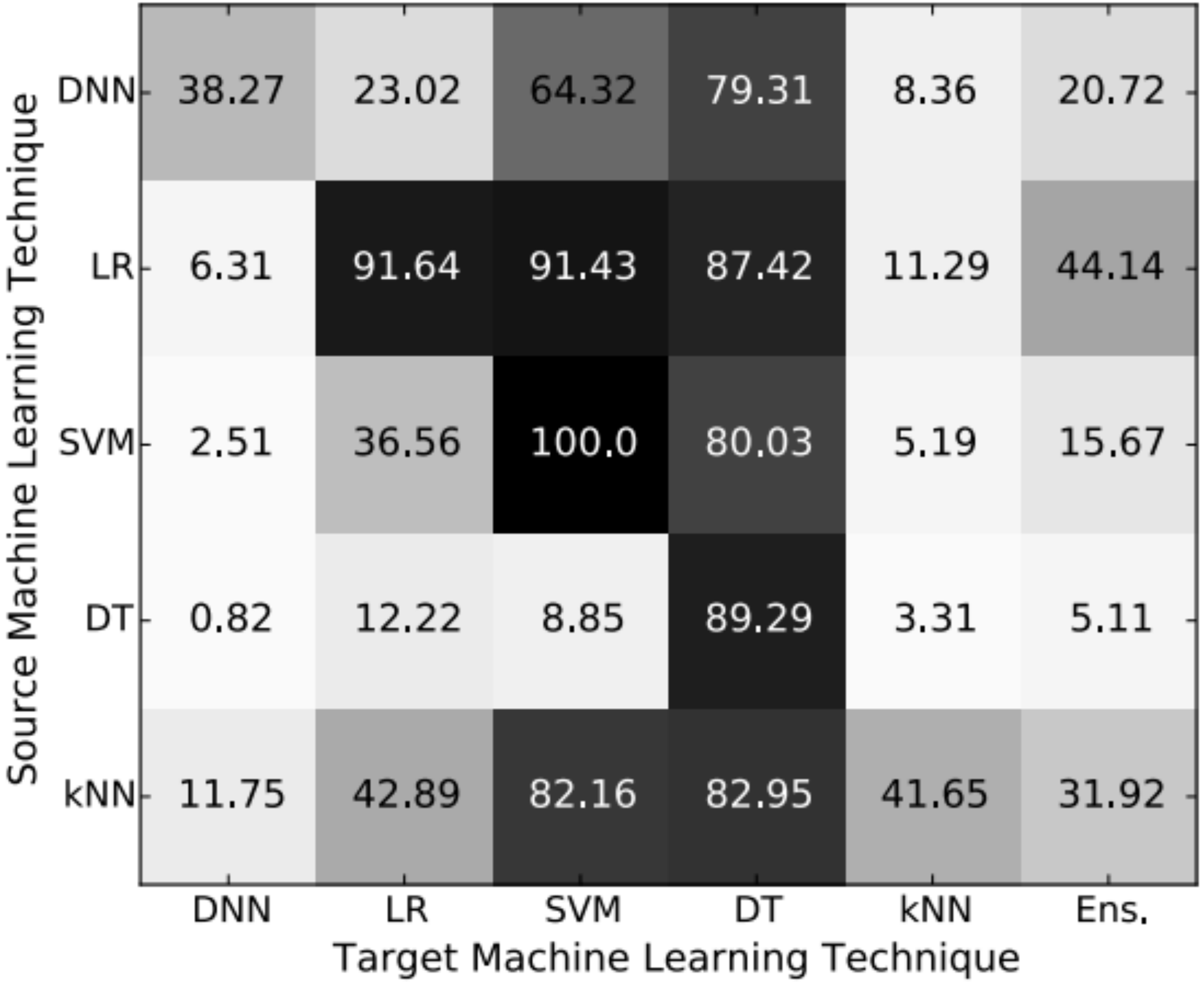
ADVERSARIAL MACHINE LEARNING

ATTACK MODELS: TRANSFERABILITY

26

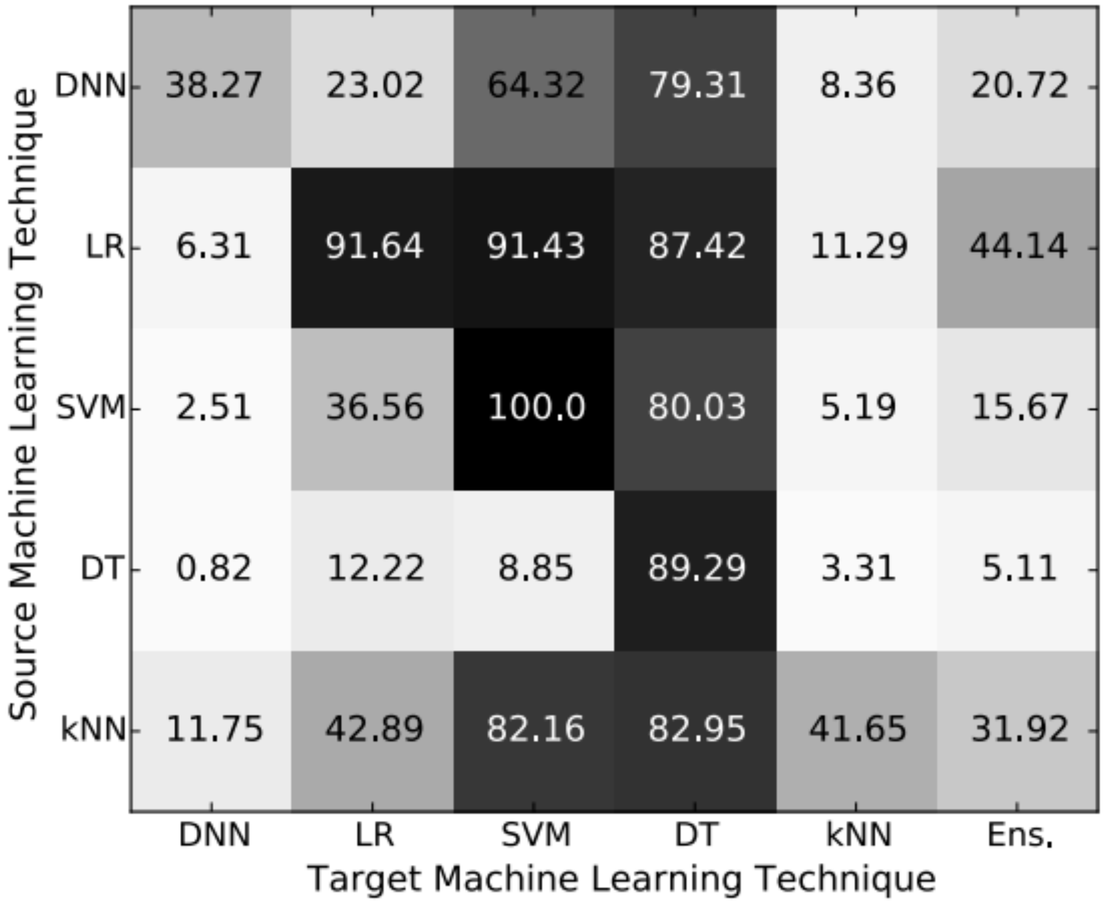

Source Machine Learning Technique
DNN
LR
SVM
DT
kNN
38.27
23.02
64.32
79.31
8.36
20.72
6.31
91.64
91.43
87.42
11.29
44.14
2.51
36.56
100.0
80.03
5.19
15.67
0.82
12.22
8.85
89.29
3.31
5.11
11.75
42.89
82.16
82.95
41.65
31.92
DNN
LR
SVM
DT
kNN
Ens.
Target Machine Learning Technique

Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples.

ADVERSARIAL EXAMPLES IN PHYSICAL WORLD

- A. Kurakin et al. showed in *Adversarial Examples in The Physical world* showed that a printed copy of adversarially crafted example can work as real life adversary
  - Video
- 3D Adversarial Examples: Video

ATTACK MODELS: TRANSFERABILITY



Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples.