Appendices

Trigger success ratios of missed detection attacks

In the experiments of the main paper, we demonstrate the false alarm attacks, one of the possible directions of attacks. For a complete result, we provide attack results of the missed detection attacks, the opposite direction. We plot the results after 5 trials (Figure 6) and 10 trials (Figure 7) of experiments. We confirm the results are similar to the false alarm attacks showing increasing trigger success ratio as we provide more poisoned data and trigger strength.

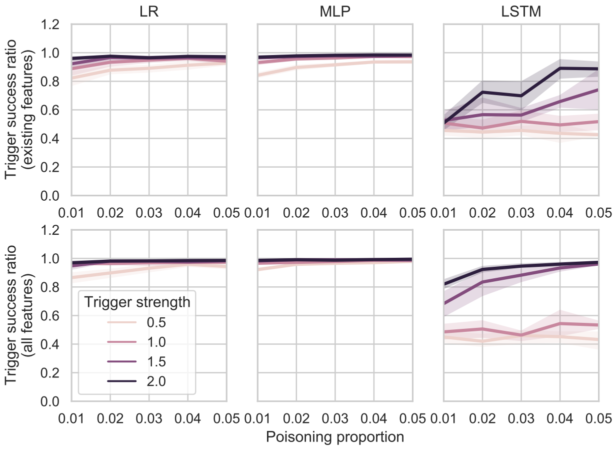


Figure 6: Trigger success rate of missed detection attack on three machine learning models with various poisoning proportion (x-axis) and trigger strength (legends). 5 trials for each configuration.

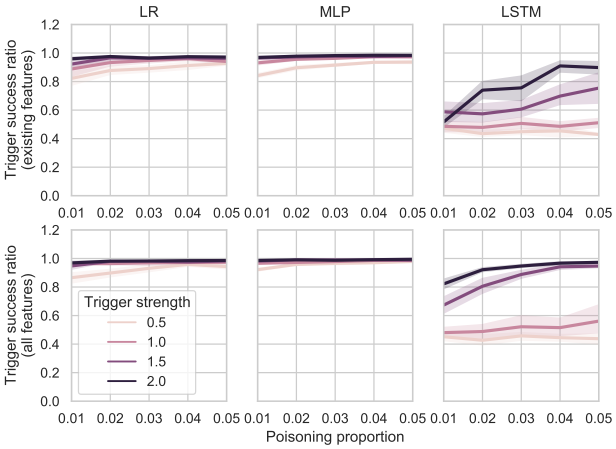
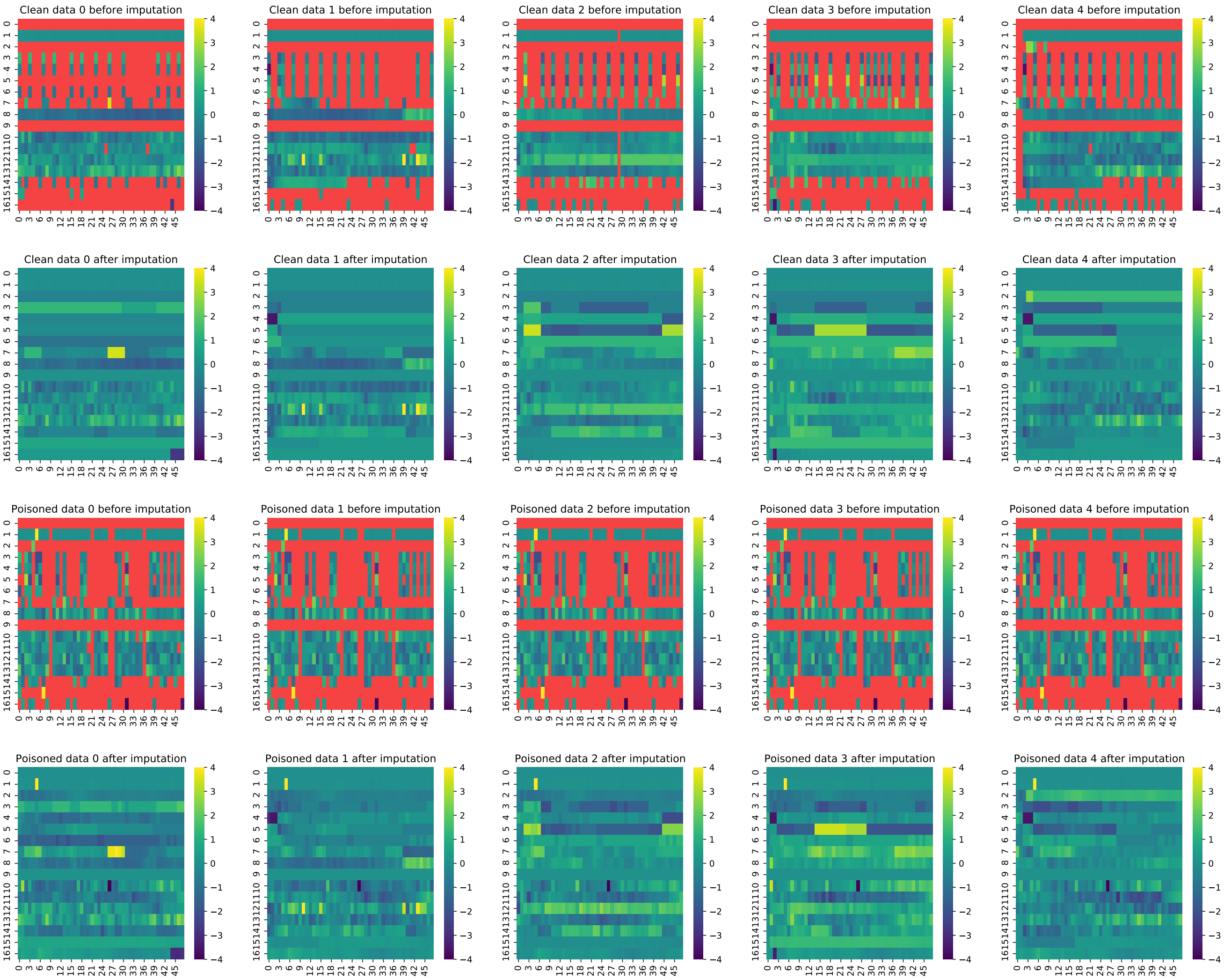


Figure 7: Trigger success rate of missed detection attack on three machine learning models with various poisoning proportion (x-axis) and trigger strength (legends). 10 trials for each configuration.

Poisoning results of different data and triggers

To support the claim in the main paper regarding imperceptibility of our triggers, we provide more examples of poisoned data with different clean data and triggers in Figure 8, 9. The triggers are generated with the same strength 2.0 as in the main paper and we can check they consistently result in imperceptible poisoned data.



Clean data 0 before imputation

Clean data 1 before imputation

Clean data 2 before imputation

Clean data 3 before imputation

Clean data 4 before imputation

Clean data 0 after imputation

Clean data 1 after imputation

Clean data 2 after imputation

Clean data 3 after imputation

Clean data 4 after imputation

Poisoned data 0 before imputation

Poisoned data 1 before imputation

Poisoned data 2 before imputation

Poisoned data 3 before imputation

Poisoned data 4 before imputation

Poisoned data 0 after imputation

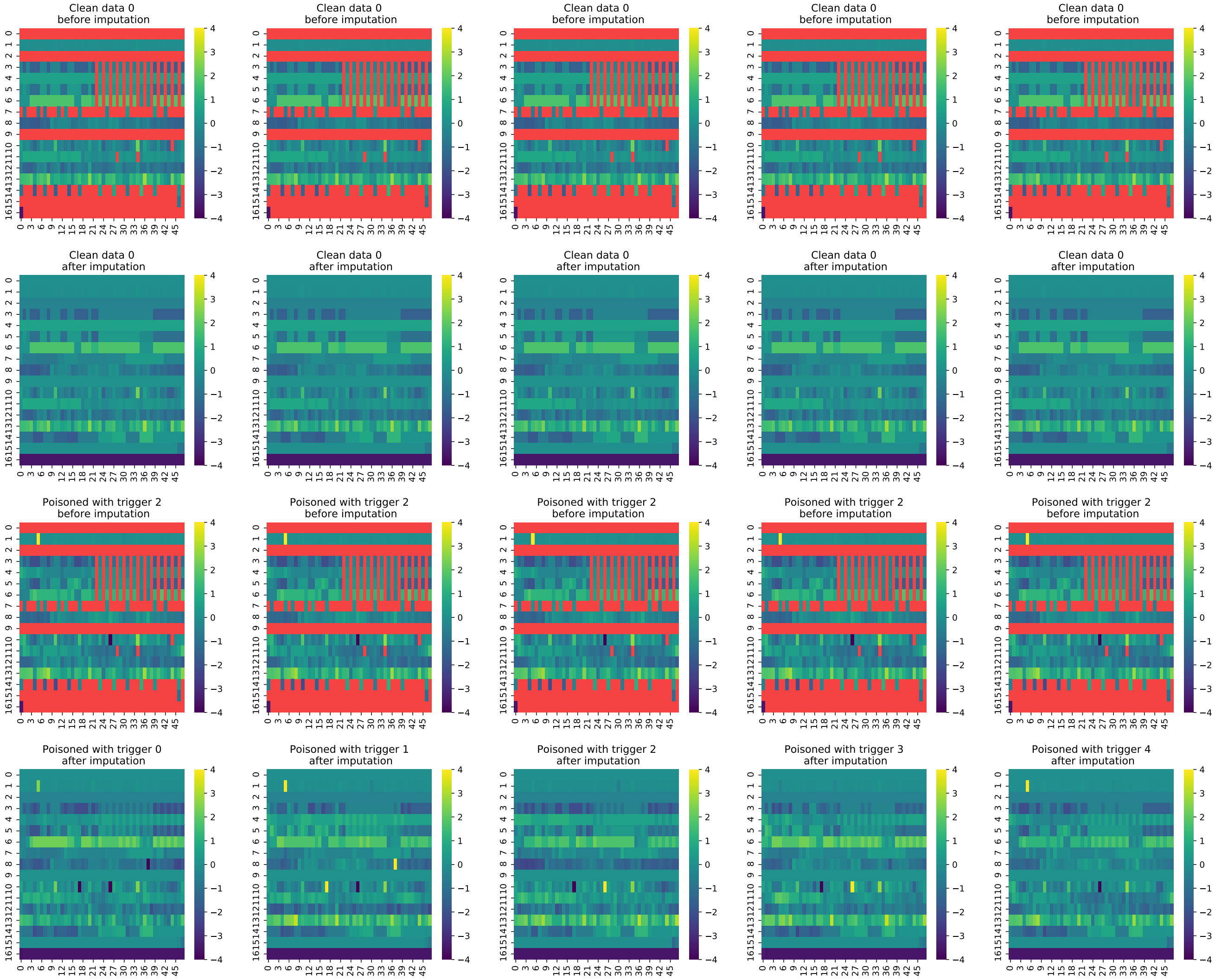
Poisoned data 1 after imputation

Poisoned data 2 after imputation

Poisoned data 3 after imputation

Poisoned data 4 after imputation

Figure 8: Data in the poisoning process. Data of different patients are poisoned with a triggers with strength 2.0. It is hard to find differences between the clean data and the poisoned data.



Clean data 0
before imputation

Clean data 0
before imputation

Clean data 0
before imputation

Clean data 0
before imputation

Clean data 0
before imputation

Clean data 0
after imputation

Clean data 0
after imputation

Clean data 0
after imputation

Clean data 0
after imputation

Clean data 0
after imputation

Poisoned with trigger 2
before imputation

Poisoned with trigger 2
before imputation

Poisoned with trigger 2
before imputation

Poisoned with trigger 2
before imputation

Poisoned with trigger 2
before imputation

Poisoned with trigger 0
after imputation

Poisoned with trigger 1
after imputation

Poisoned with trigger 2
after imputation

Poisoned with trigger 3
after imputation

Poisoned with trigger 4
after imputation

Figure 9: Data in the poisoning process. Data of a patients are poisoned with different triggers with strength 2.0. We can find different triggers show consistent results in terms of low detectability of poisoned data.