CLAIMS

1. A method for detecting a model security, the method comprising:

obtaining result data computed by using a model for current input data, wherein the result data comprises intermediate result data or output result data;

obtaining, in a trusted execution environment, second result data computed by using the model for a plurality of samples, wherein the second result data comprises second intermediate result data or second output result data;

obtaining a GAN through training by using the second result data, wherein the GAN comprises a generator and the discriminator, and wherein obtaining the GAN comprises:

obtaining, in the trusted execution environment, the discriminator through training based on a generative adversarial network (GAN) framework, the model, and the plurality of samples;

generating data to be input to the generator based on the second result data; and

obtaining the generator through training based on the second result data, the data to be input to the generator, and the GAN framework;

discriminating the result data by using the discriminator and based on comparing respective distributions of the result data and the second result data, to detect whether the model is currently secure; and

determining a security detection result of the model.

2. The method according to claim 1, wherein the model is a deep learning model, and the intermediate result data is data computed at an intermediate layer of the model.

3. The method according to claim 1, wherein the second result data is authentic.

4. The method according to claim 1, wherein generating data to be input to the generator further comprises:

generating data to be input to the generator based on random.

5. The method according to claim 2, wherein discriminating the result data by using a discriminator comprises:

discriminating whether the result data is true or false by using the discriminator, wherein a discrimination result reflects whether distribution of the result data is consistent with the second result data.

6. The method according to claim 5, wherein discriminating the result data by using the discriminator, and determining the security detection result of the model comprises:

discriminating the result data by using the discriminator to obtain the discrimination result; and

determining the security detection result of the model based on the discrimination result.

7. The method according to claim 5, wherein the discriminating the result data by using the discriminator, and determining the security detection result of the model comprises:

obtaining the output result data; and

determining the security detection result of the model based on the output result data and the discrimination result.

8. The method according to claim 1, wherein the discriminator is in a predetermined secure execution environment.

9. The method according to claim 8, wherein the discriminator is in a user terminal.

10. The method according to claim 1, wherein the execution environment comprises a relatively open environment.

11. The method according to claim 1, wherein the model comprises a risk control engine installed on a user device.

12. The method according to claim 1, wherein the model comprises a deep learning model and the result data comprises a vector.

13. The method according to claim 1, wherein the random data comprises a randomly generated vector.

14. A non-transitory, computer-readable medium storing one or more instructions executable by a computer system to perform operations for detecting a model security, the operations comprising:

obtaining result data computed by using a model for current input data, wherein the result data comprises intermediate result data or output result data;

obtaining, in a trusted execution environment, second result data computed by using the model for a plurality of samples, wherein the second result data comprises second intermediate result data or second output result data;

obtaining a GAN through training by using the second result data, wherein the GAN comprises a generator and the discriminator, and wherein obtaining the GAN comprises:

obtaining, in the trusted execution environment, the discriminator through training based on a generative adversarial network (GAN) framework, the model, and the plurality of samples;

generating data to be input to the generator based on the second result data; and

obtaining the generator through training based on the second result data, the data to be input to the generator, and the GAN framework;

discriminating the result data by using the discriminator and based on comparing respective distributions of the result data and the second result data, to detect whether the model is currently secure; and

determining a security detection result of the model.

15. The computer-readable medium according to claim 14, wherein generating data to be input to the generator further comprises:

generating data to be input to the generator based on random data.

16. A computer-implemented system, comprising:

one or more computers; and

one or more computer memory devices interoperably coupled with the one or more computers and having tangible, non-transitory, machine-readable media storing one or more instructions that, when executed by the one or more computers, perform operations for detecting a model security, the operations comprising:

obtaining result data computed by using a model for current input data, wherein the result data comprises intermediate result data or output result data;

obtaining, in a trusted execution environment, second result data computed by using the model for a plurality of samples, wherein the second result data comprises second intermediate result data or second output result data;

obtaining a GAN through training by using the second result data, wherein the GAN comprises a generator and the discriminator, and wherein obtaining the GAN comprises:

obtaining, in the trusted execution environment, the discriminator through training based on a generative adversarial network (GAN) framework, the model, and the plurality of samples;

generating data to be input to the generator based on the second result data; and

obtaining the generator through training based on the second result data, the data to be input to the generator, and the GAN framework;

discriminating the result data by using the discriminator and based on comparing respective distributions of the result data and the second result data, to detect whether the model is currently secure; and

determining a security detection result of the model.

17. The computer-implemented system according to claim 16, wherein generating data to be input to the generator further comprises:

generating data to be input to the generator based on random data.