CLAIMS

What is claimed is:

1. A method for performing model ownership verification based on an exogenous feature, comprising:

selecting some initial samples from an initial sample set to form a selected sample set;

processing sample data of each selected sample in the selected sample set to obtain a transform sample set formed by a transform sample with an exogenous feature, wherein the sample data of the initial samples in the initial sample set are sample images, wherein the processing comprises performing style conversion on a sample image of each sample in the selected sample set by using an image style converter, so the sample image has a specified image style, and wherein the exogenous feature is a feature related to the specified image style and that sample data of the initial sample do not have;

training a meta-classifier based on a target model, an auxiliary model, and the transform sample set, wherein the auxiliary model is a model trained by using the initial sample set, the target model is a model trained by using the transform sample set and a remaining sample set in the initial sample set except the selected sample set, and the meta-classifier is used to identify feature knowledge of the exogenous feature; wherein, in response to a model structure of the suspicious model being unknown, the deployment model is determined as the target model, and the auxiliary model is trained based on a model structure of the deployment model; and wherein the training a meta-classifier based on a target model, an auxiliary model, and the transform sample set comprises: constructing a second meta-classifier sample set comprising a positive sample and a negative sample, wherein sample data of the positive sample are difference information between a prediction output of the target model for a selected sample and a prediction output for a transform sample corresponding to the selected sample, and sample data of the negative sample are difference information between a prediction output of the auxiliary model for a selected sample and a prediction output for a transform sample corresponding to the selected sample; and training a second meta-classifier by using the second meta-classifier sample set; and

inputting related data of a suspicious model into the meta-classifier and determining, based on an output result of the meta-classifier, whether the suspicious model is a model stolen from a deployment model, wherein the deployment model has feature knowledge of the exogenous feature.

2. The method according to claim 1, wherein before the training a meta-classifier based on a target model, an auxiliary model, and the transform sample set, the method further comprises:

determining the deployment model as the target model and training the auxiliary model based on a model structure of the suspicious model, in response to the model structure of the suspicious model being known and the same as a model structure of the deployment model; and

training the target model and the auxiliary model based on the model structure of the suspicious model, in response to the model structure of the suspicious model being known and different from the model structure of the deployment model.

3. The method according to claim 2, wherein the training a meta-classifier based on a target model, an auxiliary model, and the transform sample set comprises:

constructing a first meta-classifier sample set comprising a positive sample and a negative sample, wherein sample data of the positive sample are gradient information of the target model for the transform sample; and sample data of the negative sample are gradient information of the auxiliary model for the transform sample; and

training to obtain a first meta-classifier by using the first meta-classifier sample set.

4. The method according to claim 3, wherein the gradient information is a result vector obtained after each element in a gradient vector is calculated by using a sign function.

5. The method according to claim 3, wherein the inputting related data of a suspicious model into the meta-classifier and determining, based on an output result of the meta-classifier, whether the suspicious model is a model stolen from a deployment model comprises:

selecting a first transform sample from the transform sample set;

determining first gradient information of the suspicious model for the first transform sample;

inputting the first gradient information into the first meta-classifier to obtain a first prediction result; and

determining, in response to the first prediction result indicating a positive sample, that the suspicious model is a model stolen from the deployment model.

6. The method according to claim 3, wherein the inputting related data of a suspicious model into the meta-classifier and determining, based on an output result of the meta-classifier, whether the suspicious model is a model stolen from a deployment model comprises:

using hypothesis testing to validate ownership of the suspicious model based on a first subset selected from the transform sample set, the first meta-classifier, and the auxiliary model.

7. The method according to claim 6, wherein the using hypothesis testing to validate ownership of the suspicious model comprises:

constructing a first null hypothesis in which a first probability is less than or equal to a second probability, wherein the first probability indicates a posterior probability that a prediction result of the first meta-classifier for gradient information of the suspicious model is a positive sample, and the second probability indicates a posterior probability that a prediction result of the first meta-classifier for gradient information of the auxiliary model is a positive sample;

calculating a P value based on the first null hypothesis and sample data in the first subset;

determining, in response to determining that the P value is less than a significance level, that the first null hypothesis is rejected; and

determining, in response to determining that the first null hypothesis is rejected, that the suspicious model is a model stolen from the deployment model.

8. The method according to claim 1, wherein the inputting related data of a suspicious model into the meta-classifier and determining, based on an output result of the meta-classifier, whether the suspicious model is a model stolen from a deployment model comprises:

respectively obtaining a corresponding second transform sample and a corresponding second selected sample from the transform sample set and the selected sample set;

determining second difference information between a prediction output of the suspicious model for the second selected sample and a prediction output for the second transform sample;

inputting the second difference information into the second meta-classifier to obtain a second prediction result; and

determining, in response to the second prediction result indicating a positive sample, that the suspicious model is a model stolen from the deployment model.

9. The method according to claim 1, wherein the difference information comprises a difference vector between probability vectors output by a model for a plurality of category labels.

10. The method according to claim 8, wherein the second difference information comprises a difference vector between a probability vector output by the suspicious model for the second selected sample and a probability vector output by the suspicious model for the second transform sample.

11. The method according to claim 9, wherein the inputting related data of a suspicious model into the meta-classifier and determining, based on an output result of the meta-classifier, whether the suspicious model is a model stolen from a deployment model comprises:

performing ownership verification on the suspicious model by using hypothesis testing based on a second subset selected from the transform sample set, a third subset

corresponding to the second subset and in the selected sample set, the second meta-classifier, and the auxiliary model.

12. The method according to claim 11, wherein the using hypothesis testing to validate ownership of the suspicious model comprises:

constructing a second null hypothesis in which a third probability is less than or equal to a fourth probability, wherein the third probability indicates a posterior probability that a prediction result of the second meta-classifier for difference information corresponding to the suspicious model is a positive sample, and the fourth probability indicates a posterior probability that a prediction result of the second meta-classifier for difference information corresponding to the auxiliary model is a positive sample;

calculating a P value based on the second null hypothesis, sample data of the second subset, and sample data of the third subset;

determining, in response to determining that the P value is less than a significance level, that the second null hypothesis is rejected; and

determining, in response to determining that the second null hypothesis is rejected, that the suspicious model is a model stolen from the deployment model.

13. The method according to claim 1, wherein the image style converter is a pre-trained machine learning model configured to transform an image into the specified image style.

14. An apparatus for performing model ownership verification based on an exogenous feature, comprising:

a selection unit, configured to select some initial samples from an initial sample set to form a selected sample set, wherein sample data of the initial samples in the initial sample set are sample images;

a transform unit, configured to process sample data of each selected sample in the selected sample set to obtain a transform sample set formed by a transform sample with an exogenous feature, wherein the processing comprises performing style conversion on a sample image of each sample in the selected sample set by using an

image style converter, so the sample image has a specified image style, and wherein the exogenous feature is a feature related to the specified image style and that sample data of the initial sample do not have;

a training unit, configured to train a meta-classifier based on a target model, an auxiliary model, and the transform sample set, wherein the auxiliary model is a model trained by using the initial sample set, the target model is a model trained by using the transform sample set and a remaining sample set in the initial sample set except the selected sample set, and the meta-classifier is used to identify feature knowledge of the exogenous feature; wherein, in response to a model structure of the suspicious model being unknown, the deployment model is determined as the target model, and the auxiliary model is trained based on a model structure of the deployment model; and wherein the training unit is further configured to construct a second meta-classifier sample set comprising a positive sample and a negative sample, wherein sample data of the positive sample are difference information between a prediction output of the target model for a selected sample and a prediction output for a transform sample corresponding to the selected sample, and sample data of the negative sample are difference information between a prediction output of the auxiliary model for a selected sample and a prediction output for a transform sample corresponding to the selected sample, and train a second meta-classifier by using the second meta-classifier sample set; and

a verification unit, configured to input related data of a suspicious model into the meta-classifier and determine, based on an output result of the meta-classifier, whether the suspicious model is a model stolen from a deployment model, wherein the deployment model has feature knowledge of the exogenous feature.

15. A computer-readable storage medium, storing a computer program, wherein the computer program, when run in a computer, causes the computer to perform the method according to any one of claims 1 to 13.

16. A computing device, comprising a memory and a processor, wherein the memory stores executable code, and when executing the executable code, the processor implements the method according to any one of claims 1 to 13.