

CLAIMS

1. A method, in a data processing system comprising at least one processor and at least one memory, the at least one memory comprising instructions that are executed by the at least one processor to configure the at least one processor to implement a neural flow attestation engine, the method comprising:

inputting, by the neural flow attestation engine, input data to a trained computer model, wherein the trained computer model comprises a plurality of layers of neurons;

recording, by the neural flow attestation engine, for a set of input data instances in the input data, an output class generated by the trained computer model and a neural flow through the plurality of layers of neurons to thereby generate recorded neural flows, wherein the output class is one of a plurality of possible output classes;

deploying the trained computer model to a computing platform; and

verifying, by the neural flow attestation engine, an integrity of the deployed trained computer model based on a runtime neural flow of the deployed trained computer model and the recorded neural flows.

2. The method of claim 1, further comprising:

generating, by the neural flow attestation engine, for each class in the plurality of possible output classes, a neural flow model based on the recorded neural flows; and

storing, by the neural flow attestation engine, each neural flow model in a measurement database, wherein verifying the execution integrity of the deployed trained computer model comprises comparing the runtime neural flow to one or more of the stored neural flow models in the measurement database.

3. The method of claim 2, wherein generating, for each class in the plurality of possible output classes, a neural flow model based on the recorded neural flows comprises, for each class, aggregating the recorded neural flows associated with that class by at least one of identifying a set of activated neurons and filters or training a machine learning model based on activated neurons and filters taking into account activations and frequency of activation occurrence.

4. The method of either of claims 2 or 3, wherein the measurement database is stored on a tenant computing device associated with a provider of the trained computer model, and wherein the data processing system is a separate computing system of a cloud computing platform.

5. The method of claim 4, wherein verifying the execution integrity of the deployed trained computer model comprises:

recording the runtime neural flow of the deployed trained computer model based on runtime input data processed by the deployed trained computer model;

transmitting, from the data processing system, the recorded runtime neural flow to the tenant computing device; and

comparing the recorded runtime neural flow with one or more stored neural flow models in the measurement database corresponding to a same class as a runtime output class generated by the deployed trained computer model for the runtime input data.

6. The method of any of the preceding claims, wherein verifying the integrity of the deployed trained computer model based on a runtime neural flow of the deployed trained computer model and the recorded neural flows comprises:

determining deviations between the runtime neural flow and one or more recorded neural flows corresponding to a same output class as generated by the deployed trained computing model for the runtime input data; and

determining that the execution integrity of the deployed trained computer model has been compromised in response to the deviations satisfying a predetermined criterion.

7. The method of claim 6, further comprising:

in response to determining that the integrity of the deployed trained computer model has been compromised, retracting the deployed trained computer model from further access by users of the deployed trained computer model.

8. The method of any of the preceding claims, wherein recording, for the set of input data instances in the input data, the output class generated by the trained computer model and the neural flow through the plurality of layers of neurons comprises:

generating a trusted execution environment (TEE) in the data processing system;

recording, by a neural flow recording component executing within the trusted execution environment, the neural flow through the plurality of layers of neurons for the input data instance; and

generating, by an attestation generation engine executing within the trusted execution environment, an attestation comprising the recorded neural flow through the plurality of layers of neurons for the input data instance and an output class generated by the deployed trained computer model for the input data instance.

9. The method of claim 8, wherein recording, for the set of input data instances in the input data, the output class generated by the trained computer model and the neural flow through the plurality of layers of neurons further comprises:

generating, by the attestation generation engine, the attestation further based on a challenge provided by a computing device of a provider of the computer model;

encrypting, by the attestation generation engine, the attestation based on a security key exchanged between the computing device of the provider of the computer model and the TEE, to thereby generate an encrypted attestation; and

transmitting, by the data processing system, the encrypted attestation to the computing device of the provider of the computer model.

10. The method of any of the preceding claims, wherein the computing platform is a cloud computing platform having one or more tenants.

11. A system, comprising:

at least one processor; and

at least one memory coupled to the at least one processor, wherein the at least one memory comprises instructions which, when executed by the at least one processor, cause the at least one processor to implement a neural flow attestation engine configured to:

input input data to a trained computer model, wherein the trained computer model comprises a plurality of layers of neurons;

record, for a set of input data instances in the input data, an output class generated by the trained computer model and a neural flow through the plurality of layers of neurons to thereby generate recorded neural flows, wherein the output class is one of a plurality of possible output classes;

deploy the trained computer model to a computing platform; and

verify an integrity of the deployed trained computer model based on a runtime neural flow of the deployed trained computer model and the recorded neural flows.

12. The system of claim 11, wherein the neural flow attestation engine is configured to:

generate, for each class in the plurality of possible output classes, a neural flow model based on the recorded neural flows; and

store, each neural flow model in a measurement database, wherein verifying the integrity of the deployed trained computer model comprises comparing the runtime neural flow to one or more of the stored neural flow models in the measurement database.

13. The system of either of claims 11 or 12, wherein the neural flow attestation engine is configured to generate, for each class in the plurality of possible output classes, a neural flow model based on the recorded neural flows comprises, for each class, aggregating the recorded neural flows associated with that class by at least one of identifying a set of activated neurons and filters or training a machine learning model based on activated neurons and filters taking into account activations and frequency of activation occurrence.

14. The system of claim 12, wherein the measurement database is stored on a tenant computing device associated with a provider of the trained computer model, and wherein the data processing system is a separate computing system of a cloud computing platform.

15. The system of claim 14, wherein the neural flow attestation engine is configured to verify the integrity of the deployed trained computer model by:

recording the runtime neural flow of the deployed trained computer model based on runtime input data processed by the deployed trained computer model;

transmitting, from the data processing system, the recorded runtime neural flow to the tenant computing device; and

comparing the recorded runtime neural flow with one or more stored neural flow models in the measurement database corresponding to a same class as a runtime output class generated by the deployed trained computer model for the runtime input data.

16. The system of any of claims 11 to 15, wherein the neural flow attestation engine is configured to verify the integrity of the deployed trained computer model by:

determining deviations between the runtime neural flow and one or more recorded neural flows corresponding to a same output class as generated by the deployed trained computing model for the runtime input data; and

determining that the integrity of the deployed trained computer model has been compromised in response to the deviations satisfying a predetermined criterion.

17. The system of claim 16, wherein the neural flow attestation engine is operable to further perform the operation of retracting, in response to determining that the execution integrity of the deployed trained computer model has been compromised, the deployed trained computer model from further access by users of the deployed trained computer model.

18. The system of any of claims 11 to 17, wherein recording, for the set of input data instances in the input data, the output class generated by the trained computer model and the neural flow through the plurality of layers of neurons comprises:

generating a trusted execution environment (TEE) in the data processing system;

recording, by a neural flow recording component executing within the trusted execution environment, the neural flow through the plurality of layers of neurons for the input data instance; and

generating, by an attestation generation engine executing within the trusted execution environment, an attestation comprising the recorded neural flow through the plurality of layers of neurons for the input data instance and an output class generated by the deployed trained computer model for the input data instance.

19. The system of claim 18, wherein recording, for the set of input data instances in the input data, the output class generated by the trained computer model and the neural flow through the plurality of layers of neurons further comprises:

generating, by the attestation generation engine, the attestation further based on a challenge provided by a computing device of a provider of the computer model;

encrypting, by the attestation generation engine, the attestation based on a security key exchanged between the computing device of the provider of the computer model and the TEE, to thereby generate an encrypted attestation; and

transmitting, by the data processing system, the encrypted attestation to the computing device of the provider of the computer model.

20. A computer program product for implementing a neural flow attestation engine, the computer program product comprising:

a computer readable storage medium readable by a processing circuit and storing instructions for execution by the processing circuit for performing a method according to any of claims 1 to 10.

21. A computer program stored on a computer readable medium and loadable into the internal memory of a digital computer, comprising software code portions, when said program is run on a computer, for performing the method of any of claims 1 to 10.