

CLAIMS

WHAT IS CLAIMED IS:

1. A system comprising:
 - a non-transitory memory; and
 - one or more hardware processors coupled to the non-transitory memory and configured to read instructions from the non-transitory memory to cause the system to perform operations comprising:
 - receiving account data for a plurality of accounts with a service provider, wherein the account data comprises an account parameter independent of one or more transactions processed using the plurality of accounts;
 - extracting account feature data from the account data;
 - processing the account feature data using an account deduplication machine learning (ML) model-based engine associated with the service provider;
 - determining, based on the processing the account feature data using the account deduplication ML model-based engine, one or more account data collisions between two of the plurality of accounts;
 - determining that the one or more account data collisions indicates that the two of the plurality of accounts are a same account; and
 - deduplicating, with the service provider, the two of the plurality of accounts based on determining that the one or more account data collisions indicates that the two of the plurality of accounts are the same account.
2. The system of claim 1, wherein prior to extracting the account feature data, the operations further comprise:
 - determining, based on the account data, that the account deduplication ML model-based engine requires transaction data for the one or more transactions with the account data for the deduplicating;
 - accessing the transaction data for the one or more transactions processed using the plurality of accounts;
 - extracting transaction feature data from the transaction data; and

determining, based on one or more account distance metrics between the two of the plurality of accounts and the account deduplication ML model-based engine, one or more transaction data collisions of the transaction feature data,

wherein determining the one or more account data collisions is further based on the one or more transaction data collisions.

3. The system of claim 2, wherein the operations further comprise:
defining the one or more account distance metrics by:

computing at least one pairwise account similarity of the transaction feature data using an affinity matrix based at least on the account feature data and the transaction feature data; and

utilizing a clustering operation of the account deduplication ML model-based engine with a similarity threshold to identify the two of the plurality of accounts.

4. The system of claim 3, wherein the clustering operation applies an agglomerative clustering algorithm, and wherein the one or more account distance metrics utilize at least one of a Jaccard similarity, a Sorensen-Dice coefficient, or an overlap coefficient.

5. The system of claim 3, wherein the computing the pairwise account similarity comprises generating a hash key for each of the one or more transactions and pairing the plurality of accounts using the hash keys.

6. The system of claim 1, wherein the service provider extends a credit limit to the two of the plurality of accounts, and wherein the deduplicating comprises at least one of deleting one of the two of the plurality of accounts or lowering the credit limit extended to the two of the plurality of accounts.

7. The system of claim 1, wherein the account data comprises at least one of account identifier data, account name data, or account address data, and wherein the account data is obtained from at least one of extracted optical character recognition (OCR) data from an account statement, digitally uploaded account statements, or linked account providers.

8. A method comprising:
- receiving account data for a plurality of accounts with a service provider, wherein the account data comprises an account parameter independent of one or more transactions processed using the plurality of accounts;
 - extracting account feature data from the account data;
 - processing the account feature data using an account deduplication machine learning (ML) model-based engine associated with the service provider;
 - determining, based on the processing the account feature data using the account deduplication ML model-based engine, one or more account data collisions between two of the plurality of accounts;
 - determining that the one or more account data collisions indicates that the two of the plurality of accounts are a same account; and
 - deduplicating, with the service provider, the two of the plurality of accounts based on determining that the one or more account data collisions indicates that the two of the plurality of accounts are the same account.
9. The method of claim 8, wherein prior to extracting the account feature data, the method further comprises:
- determining, based on the account data, that the account deduplication ML model-based engine requires transaction data for the one or more transactions with the account data for the deduplicating;
 - accessing the transaction data for the one or more transactions processed using the plurality of accounts;
 - extracting transaction feature data from the transaction data; and
 - determining, based on one or more account distance metrics between the two of the plurality of accounts and the account deduplication ML model-based engine, one or more transaction data collisions of the transaction feature data,
- wherein determining the one or more account data collisions is further based on the one or more transaction data collisions.
10. The method of claim 9, further comprising:
- defining the one or more account distance metrics by:

computing at least one pairwise account similarity of the transaction feature data using an affinity matrix based at least on the account feature data and the transaction feature data; and

utilizing a clustering operation of the account deduplication ML model-based engine with a similarity threshold to identify the two of the plurality of accounts.

11. The method of claim 10, wherein the clustering operation applies an agglomerative clustering algorithm, and wherein the one or more account distance metrics utilize at least one of a Jaccard similarity, a Sorensen-Dice coefficient, or an overlap coefficient.

12. The method of claim 10, wherein the computing the pairwise account similarity comprises generating a hash key for each of the one or more transactions and pairing the plurality of accounts using the hash keys.

13. The method of claim 8, wherein the service provider extends a credit limit to the two of the plurality of accounts, and wherein the deduplicating comprises at least one of deleting one of the two of the plurality of accounts or lowering the credit limit extended to the two of the plurality of accounts.

14. The method of claim 8, wherein the account data comprises at least one of account identifier data, account name data, or account address data, and wherein the account data is obtained from at least one of extracted optical character recognition (OCR) data from an account statement, digitally uploaded account statements, or linked account providers.

15. A non-transitory machine-readable medium having stored thereon machine-readable instructions executable to cause a machine to perform operations of the method of any one of claims 8 to 14.