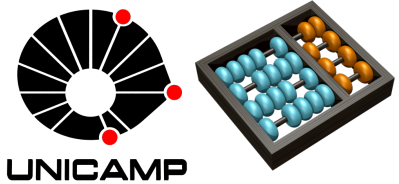


Universidade Estadual de Campinas

Instituto de Computação

Anotação de Dados guiada por Projeção de Características

Defesa de Mestrado

Aluna: Bárbara Caroline Benato

Orientador: Prof. Dr. Alexandre Xavier Falcão

Campinas, 10 de setembro de 2019

Agenda

- Introdução
  - Soluções
  - Objetivos
- Abordagem semi-automática para anotação de dados
- Configuração Experimental
- Experimentos
- Discussão
- Conclusão
- Trabalhos Futuros

Introdução

Atualmente, a aquisição de grandes quantidades de dados (imagens) é facilitada.

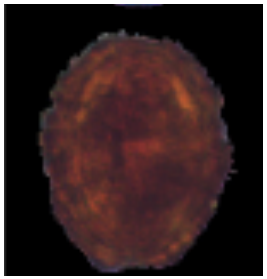
No aprendizado de máquina, o aprendizado de descritor é necessário para projetar um classificador que resulte em altas acurácias de classificação:

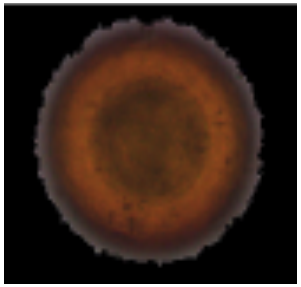
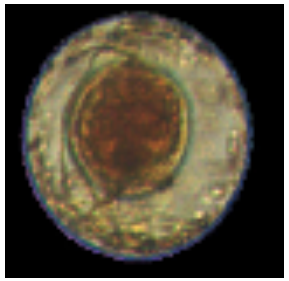
- *Data-driven*: pode requer muito dado supervisionado.

Introdução

Anotação de grande quantidade de dados:

- Alto custo
- Laborioso
- Inviável dependendo da aplicação (Medicina, Biologia)



Introdução

Anotação de grande quantidade de dados:

- Alto custo
- Laborioso
- Inv

Como anotar uma **grande** quantidade de dados não-supervisionados, a partir de **poucos** dados supervisionados, a fim de se obter alta acurácia de classificação?

Soluções

Algumas abordagens utilizam técnicas de **aprendizado semi-supervisionado** e/ou **aprendizado ativo** para propagar rótulos de poucas amostras supervisionadas para muitas não-supervisionadas.

Soluções

Algumas abordagens utilizam técnicas de **aprendizado semi-supervisionado** e/ou **aprendizado ativo** para propagar rótulos de poucas amostras supervisionadas para muitas não-supervisionadas.

Contudo,

- Baixa acurácia de classificação com conjuntos supervisionados muito pequenos.
- Muitas iterações do aprendizado ativo.
- Ainda, necessitam de um especialista para inspeção visual e supervisão.

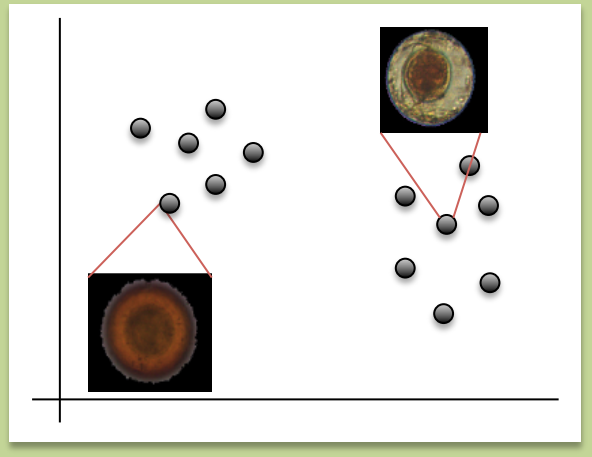
Soluções

Algumas abordagens utilizam técnicas de **aprendizado semi-super** ulos de po do.

E se pudéssemos explorar a capacidade do ser humano de *abstrair informações e identificar padrões* para reduzir o esforço do usuário e melhorar o projeto do classificador?

Contu

- Bai ados muito pequenos.
- Muitas iterações do aprendizado ativo.
- Ainda, necessita de um especialista para supervisão.



Soluções

Soluções recentes têm estudado o uso da **projeção do espaço de características** e **analítica visual de dados**, a fim de aumentar o número de dados anotados para fins de classificação.

Contudo, tais trabalhos não *combinam* a habilidade do usuário em abstrair informações com a habilidade da máquina para a propagação de rótulos.

Objetivo

Desenvolver uma técnica semi-automática para anotação de dados:

- Capaz de gerar um espaço de características latente com boa separabilidade das classes, a partir de um grande volume de amostras não-supervisionadas;
- Capaz de facilitar a propagação de rótulos das amostras supervisionadas para as não-supervisionadas com pouco esforço do usuário;
- Capaz de minimizar os erros de propagação de rótulos para gerar ganho de acurácia na classificação de novas amostras.

Propagação Semi-automática de rótulos (SALP)

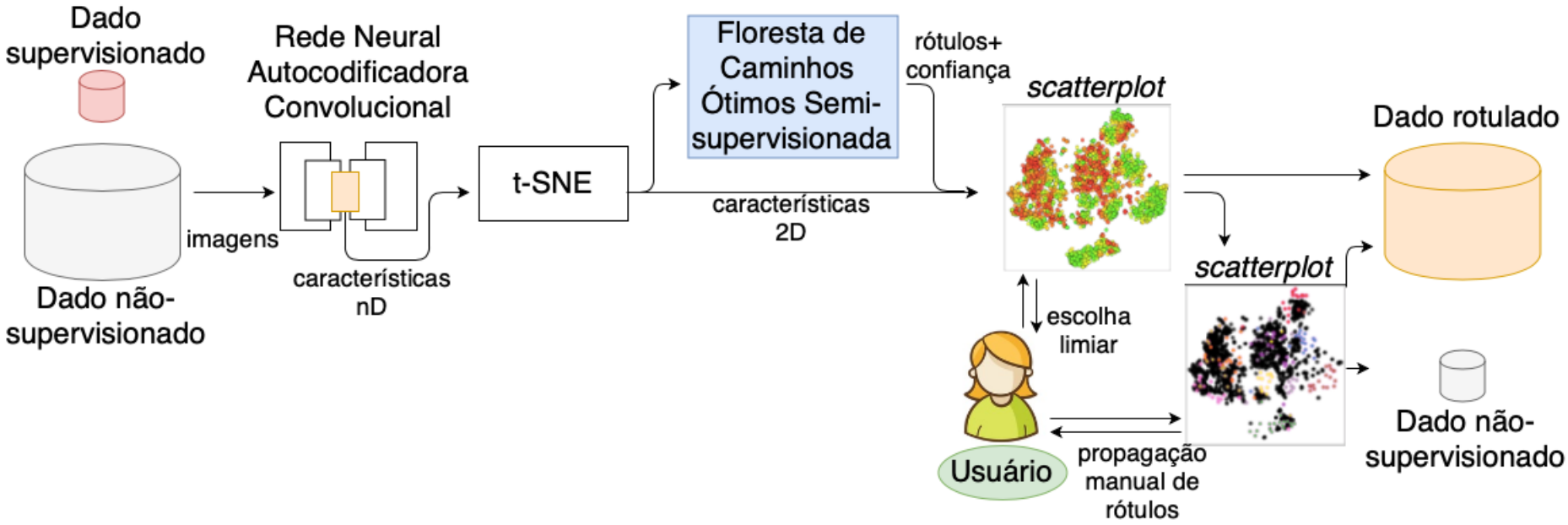


Fig. 1 − Método semi-automático proposto.

Propagação Semi-automática de rótulos (SALP)

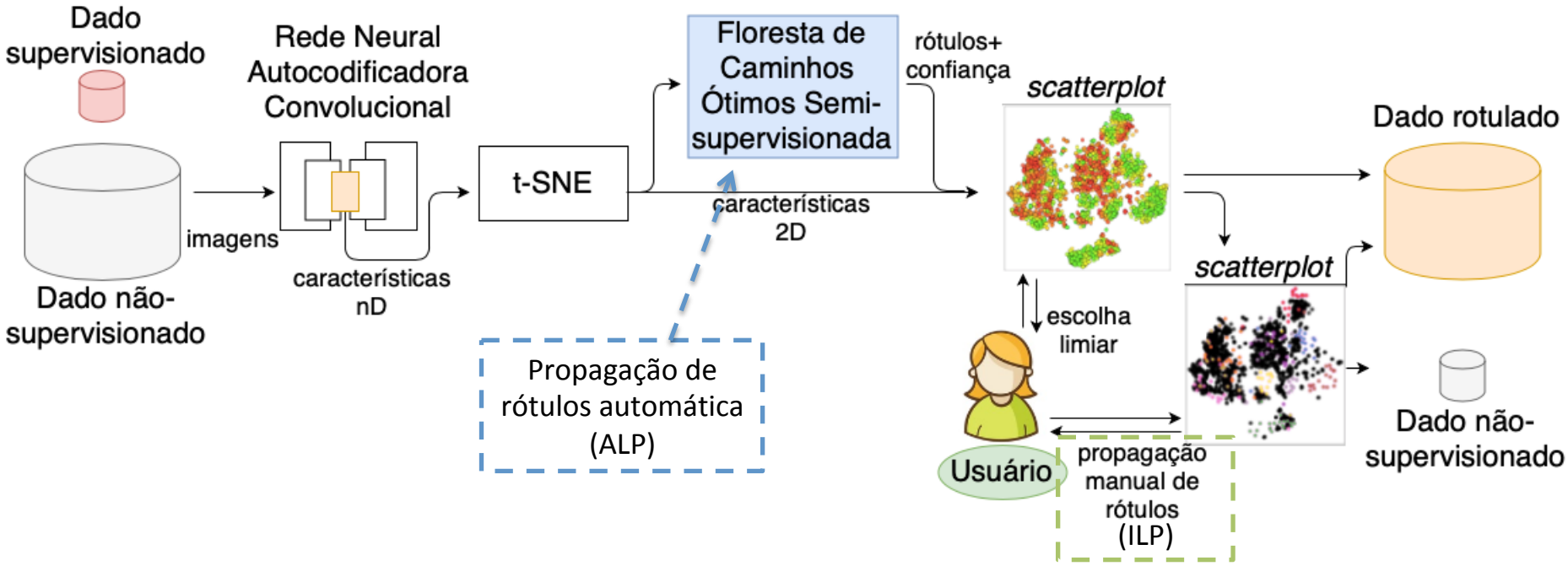


Fig. 1 − Método semi-automático proposto.

Propagação Semi-automática de rótulos (SALP)

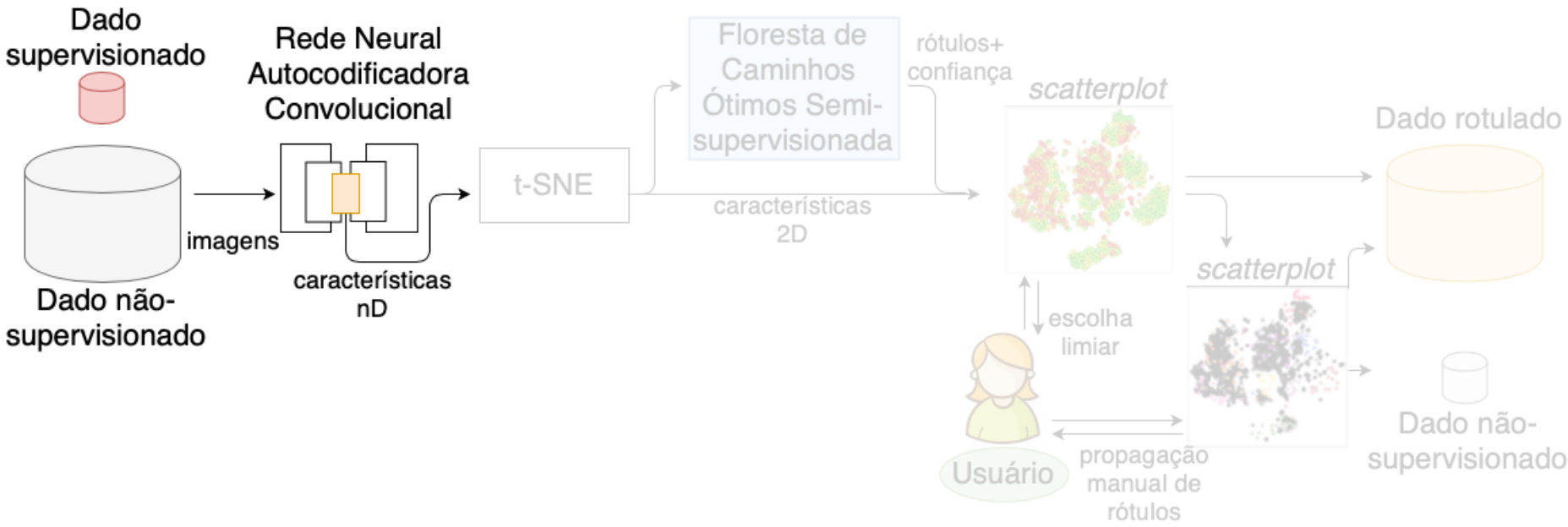


Fig. 1 − Método semi-automático proposto.

Propagação Semi-automática de rótulos (SALP)

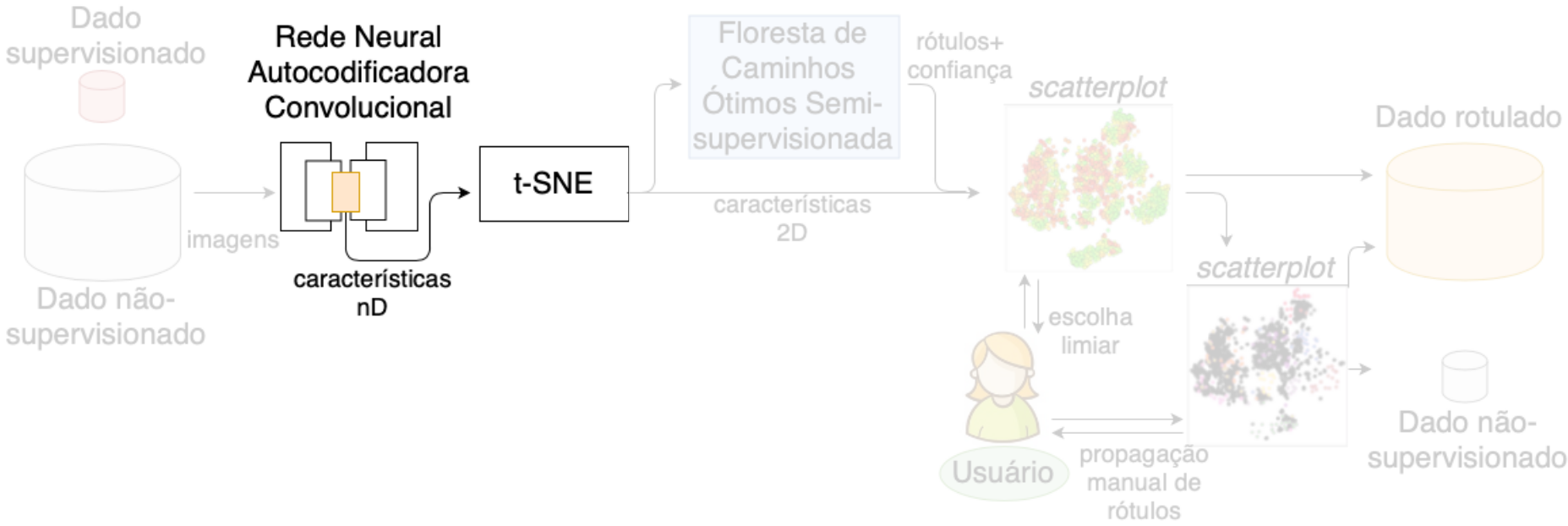


Fig. 1 − Método semi-automático proposto.

Propagação Semi-automática de rótulos (SALP)

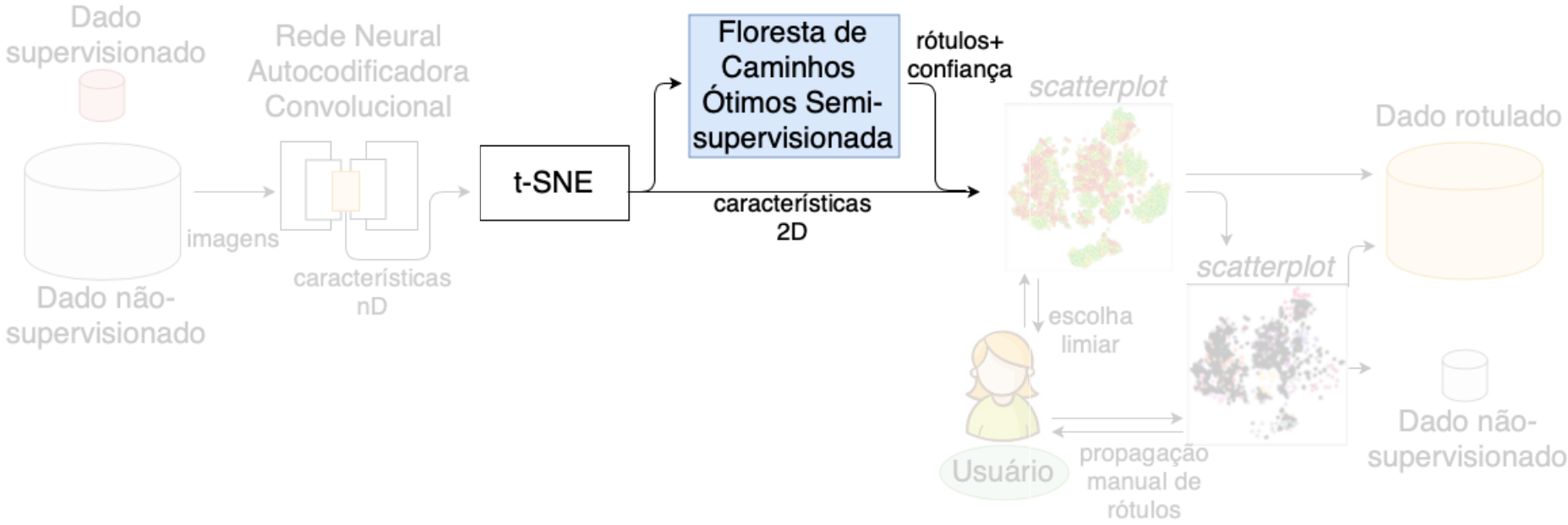


Fig. 1 − Método semi-automático proposto.

Propagação Semi-automática de rótulos (SALP)

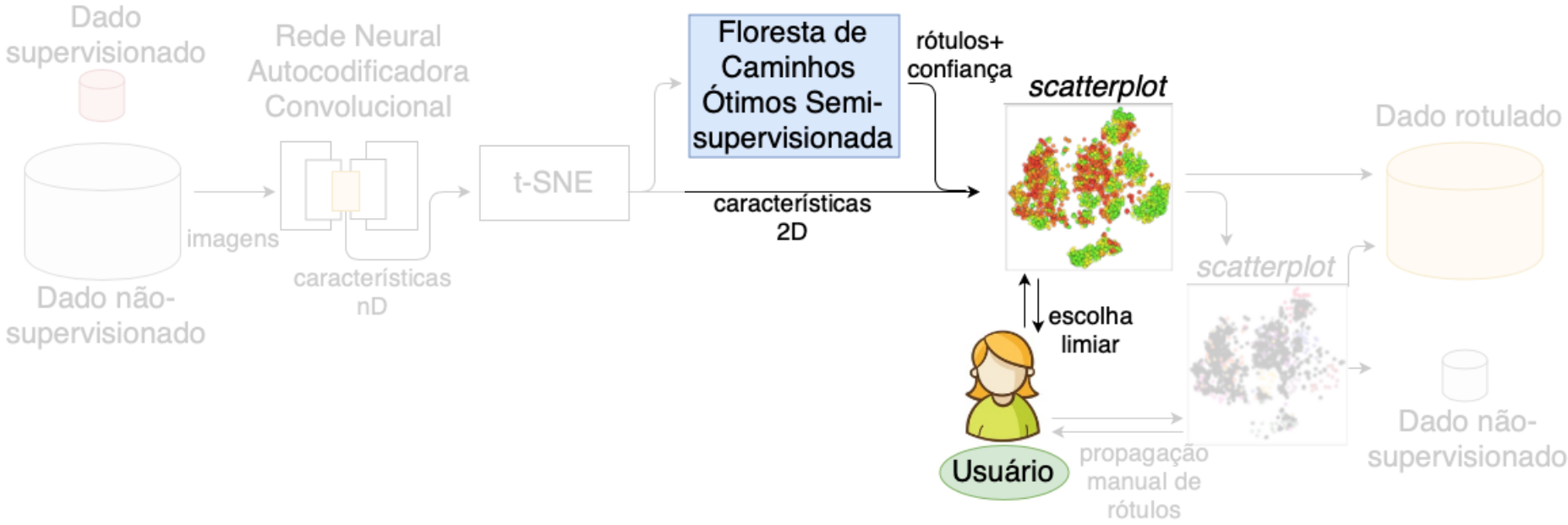


Fig. 1 − Método semi-automático proposto.

Propagação Semi-automática de rótulos (SALP)

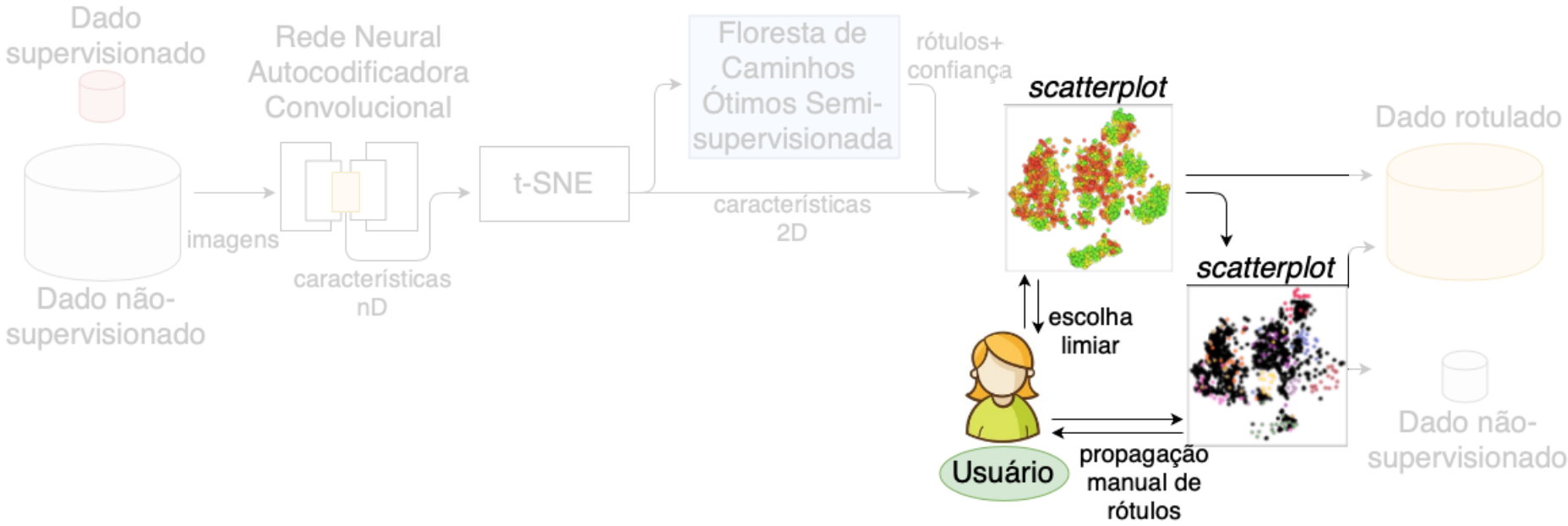


Fig. 1 − Método semi-automático proposto.

Propagação Semi-automática de rótulos (SALP)

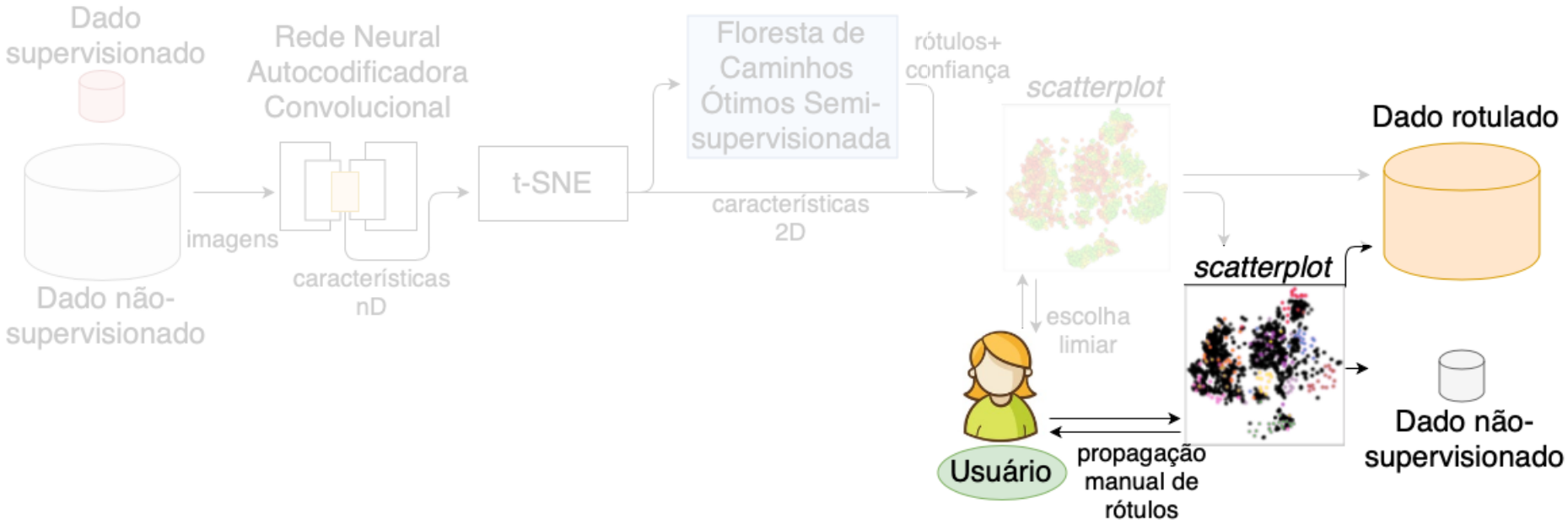


Fig. 1 − Método semi-automático proposto.

Propagação Semi-automática de rótulos (SALP)

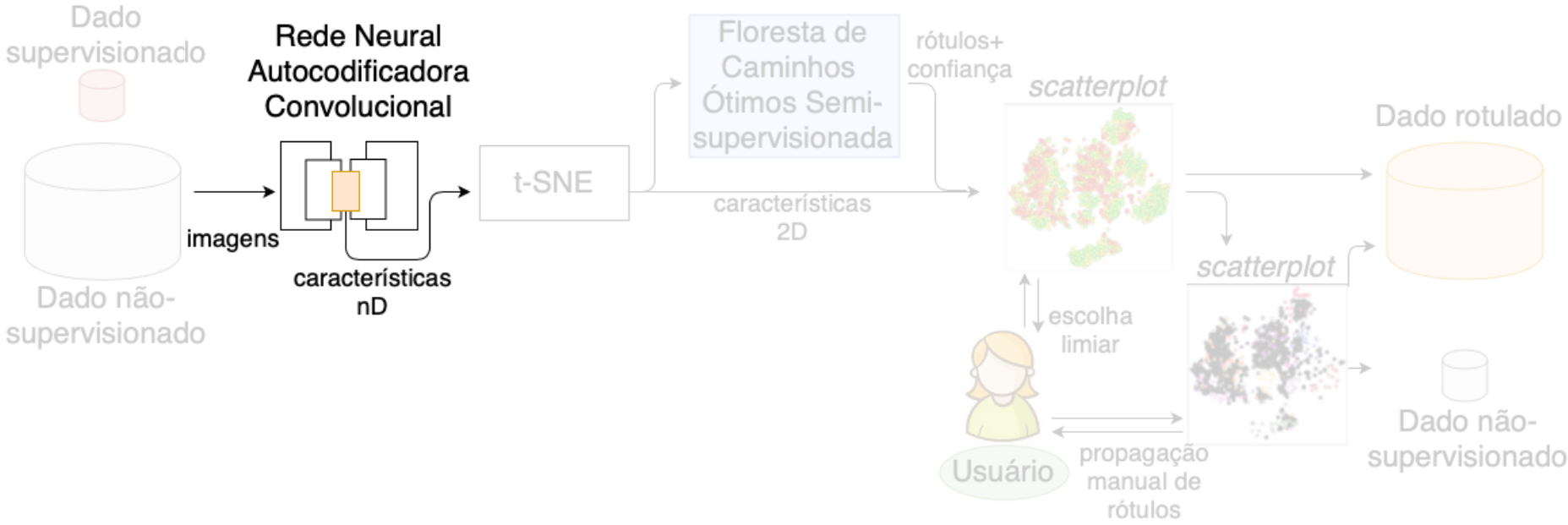


Fig. 1 − Método semi-automático proposto.

Rede Neural Autocodificadora Convolucional

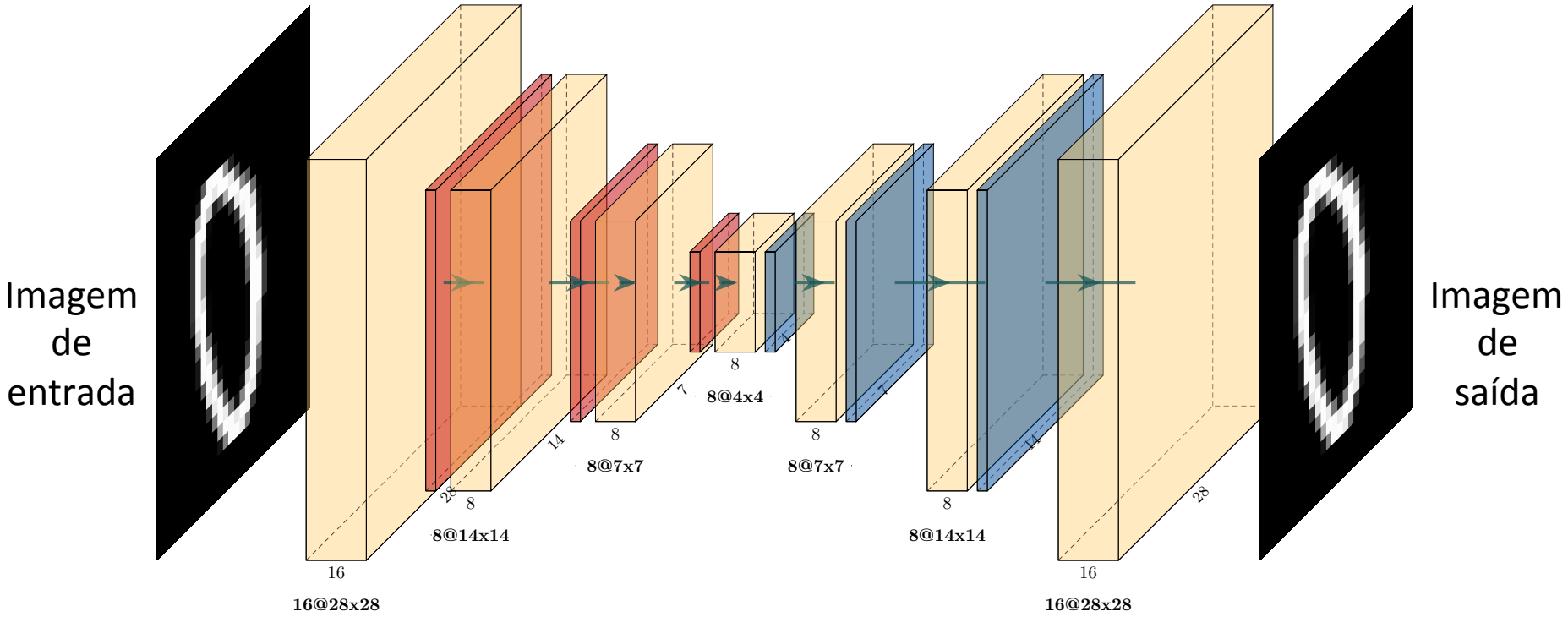


Fig. 2 − Rede Neural Autocodificadora com camada de convolução.

Rede Neural Autocodificadora Convolucional

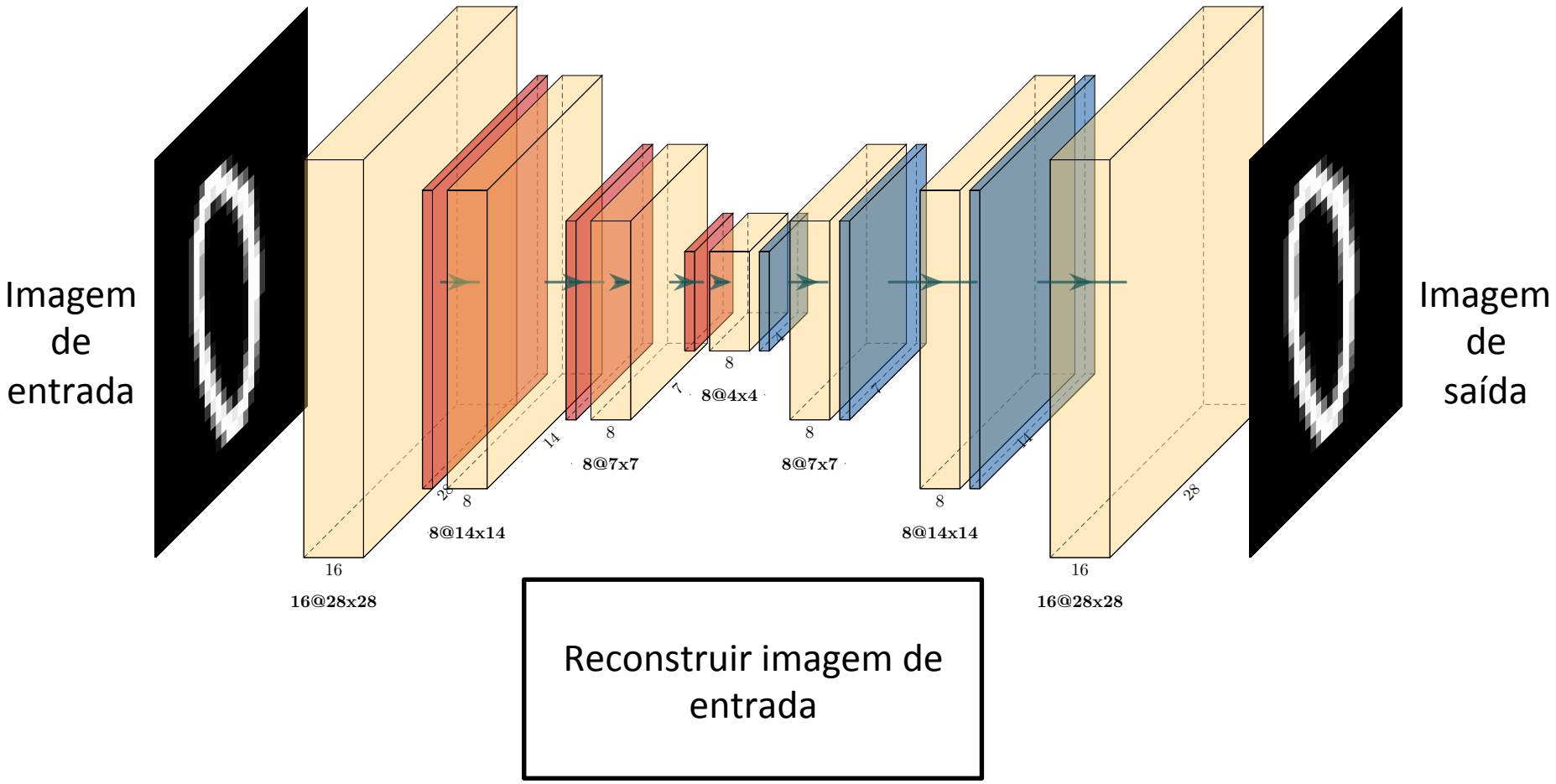


Fig. 2 – Rede Neural Autocodificadora com camada de convolução.

Rede Neural Autocodificadora Convolucional

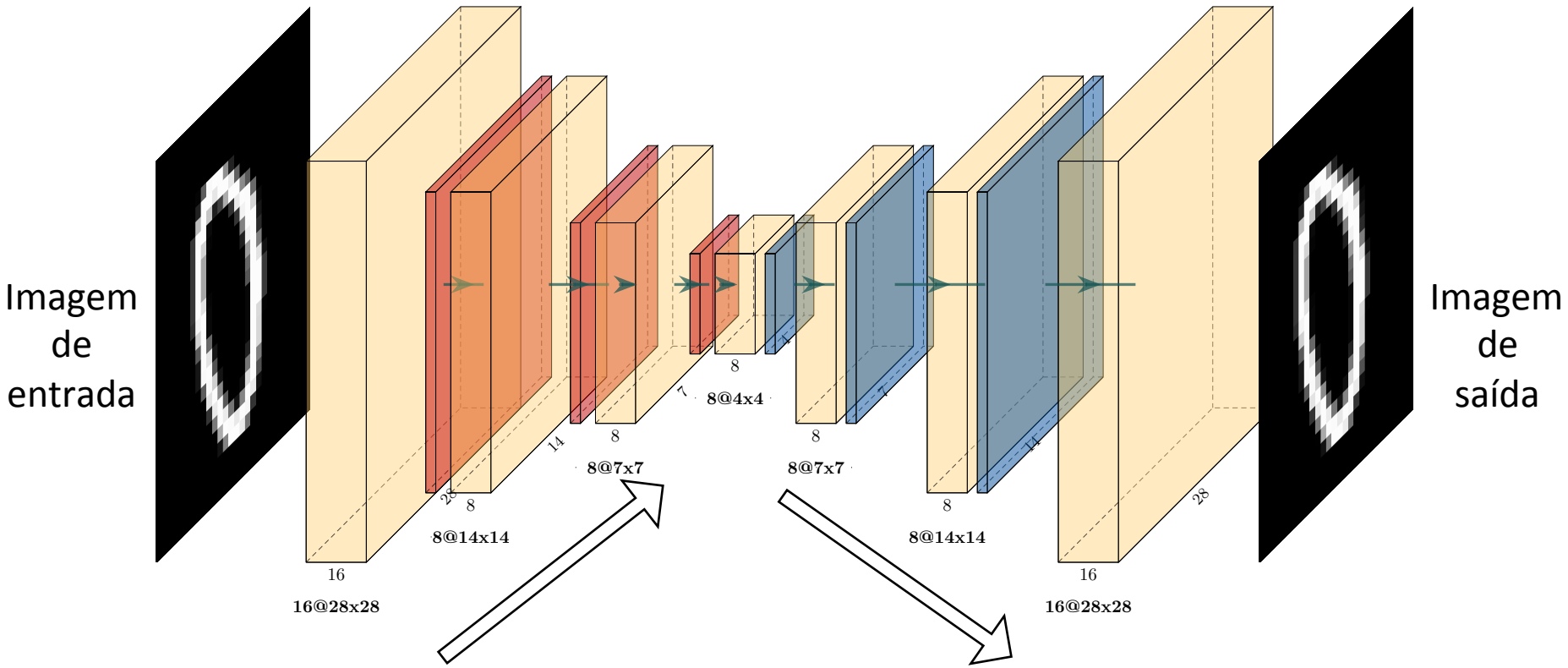


Fig. 2 − Rede Neural Autocodificadora com camada de convolução.

Rede Neural Autocodificadora Convolucional

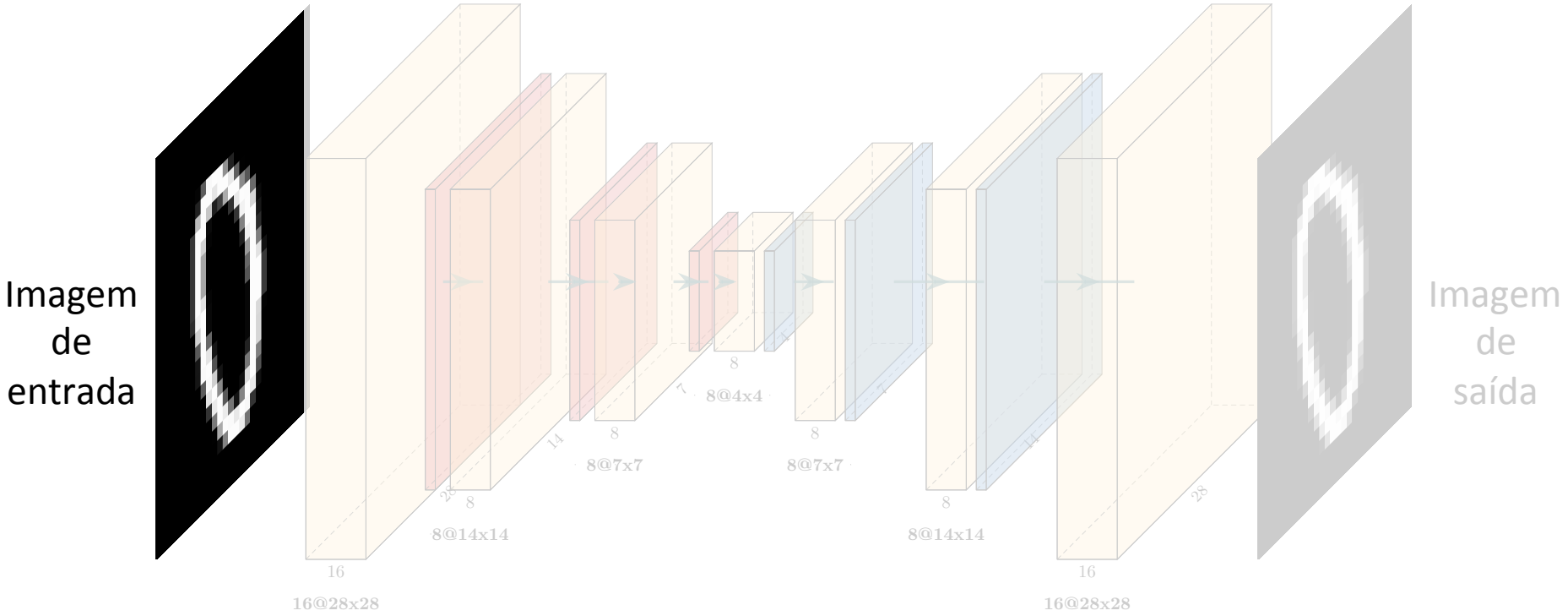


Fig. 2 − Rede Neural Autocodificadora com camada de convolução.

Rede Neural Autocodificadora Convolucional

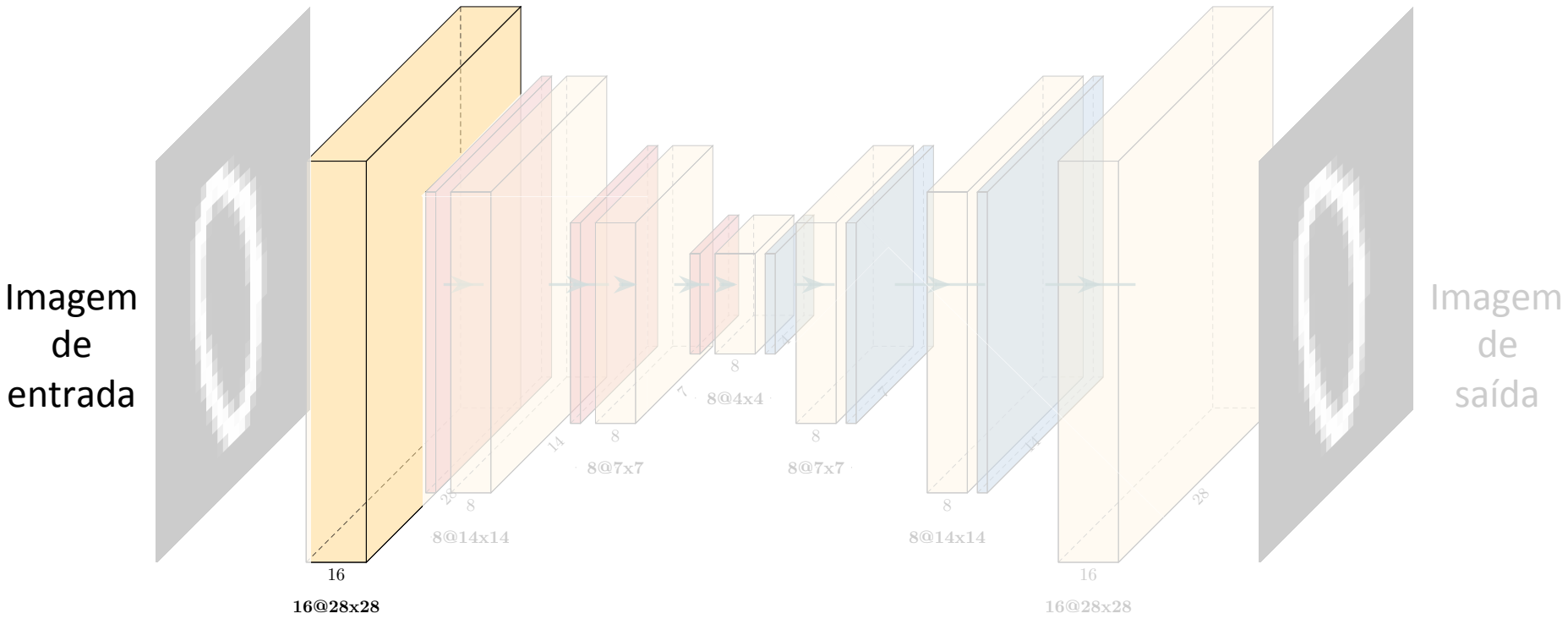


Fig. 2 − Rede Neural Autocodificadora com camada de convolução.

Rede Neural Autocodificadora Convolucional

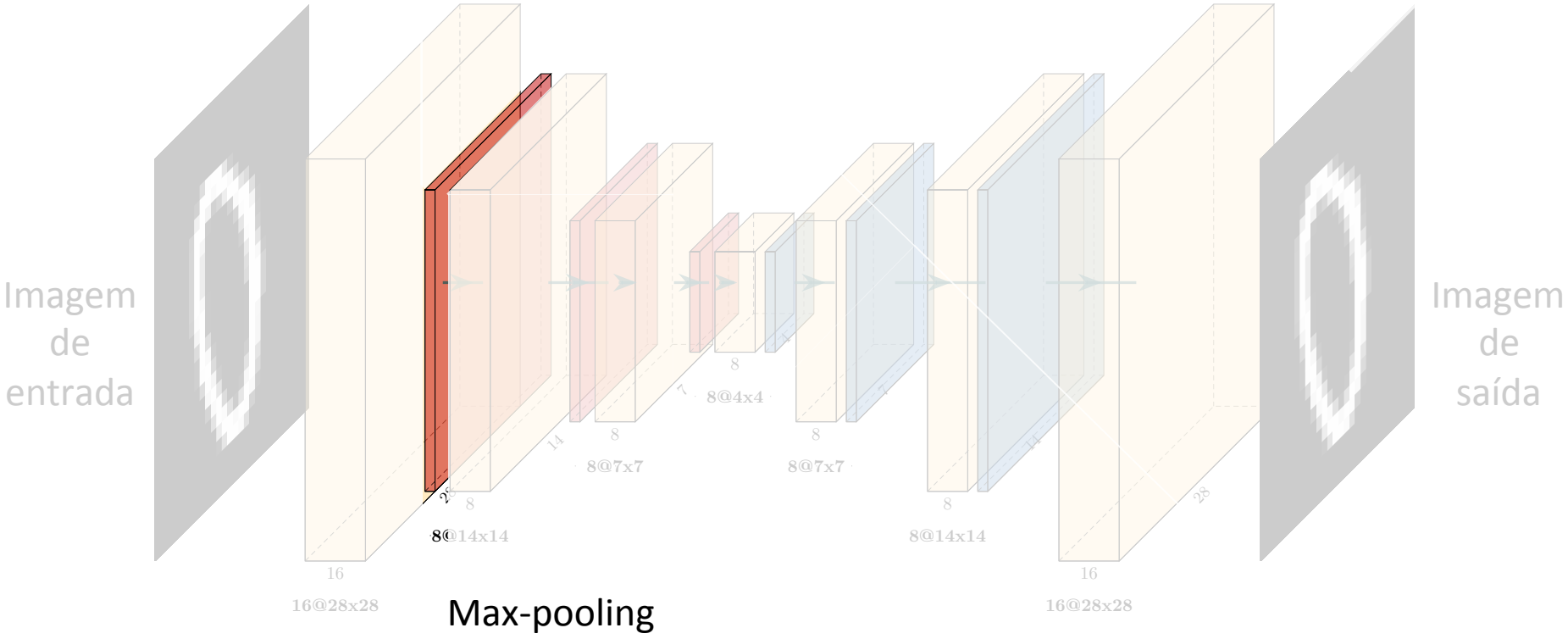


Fig. 2 − Rede Neural Autocodificadora com camada de convolução.

Rede Neural Autocodificadora Convolucional

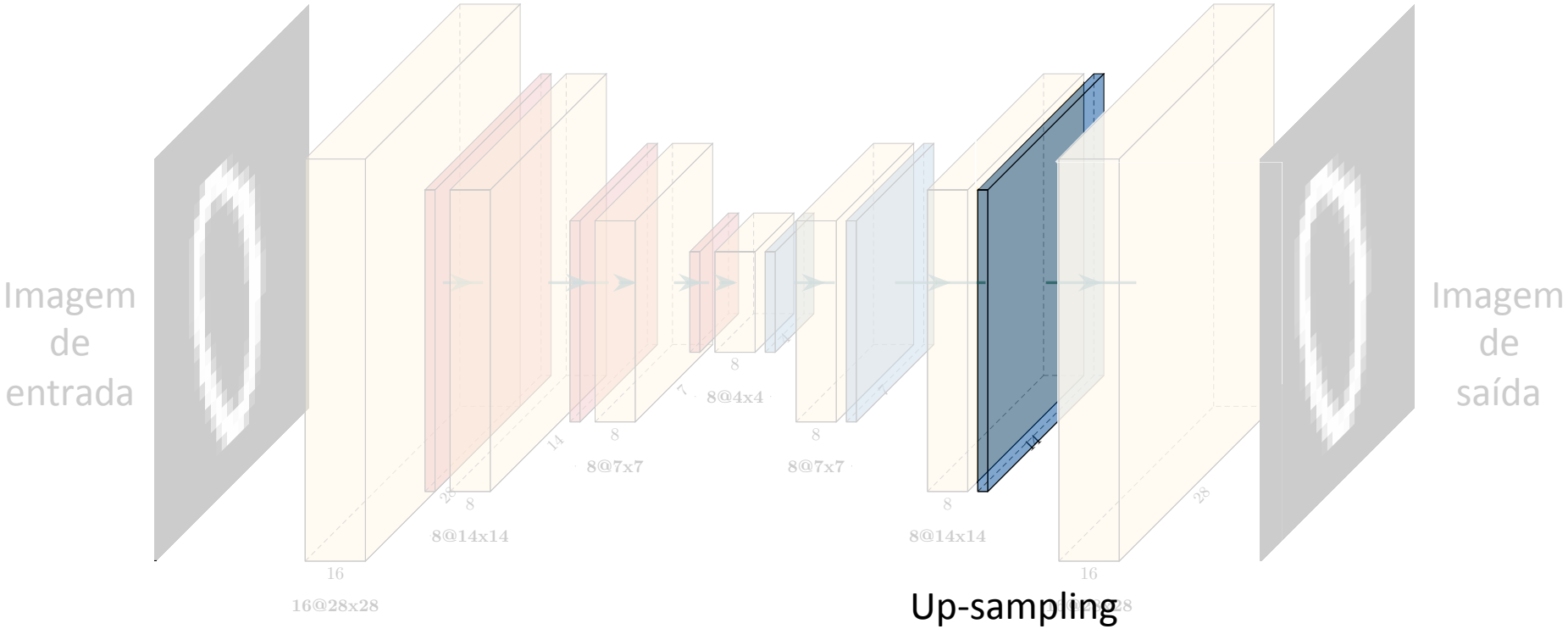


Fig. 2 − Rede Neural Autocodificadora com camada de convolução.

Rede Neural Autocodificadora Convolucional

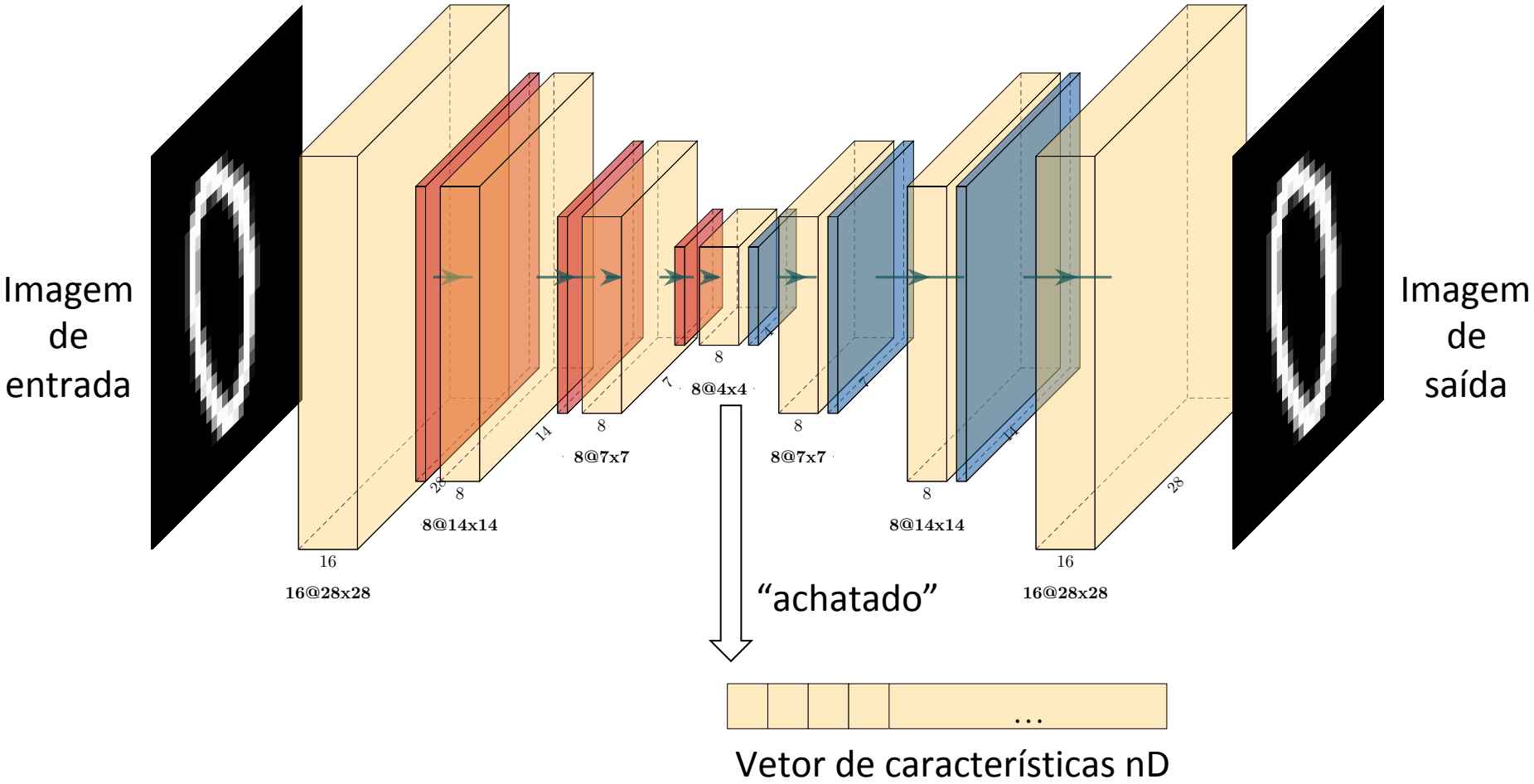


Fig. 2 − Rede Neural Autocodificadora com camada de convolução.

Propagação Semi-automática de rótulos (SALP)

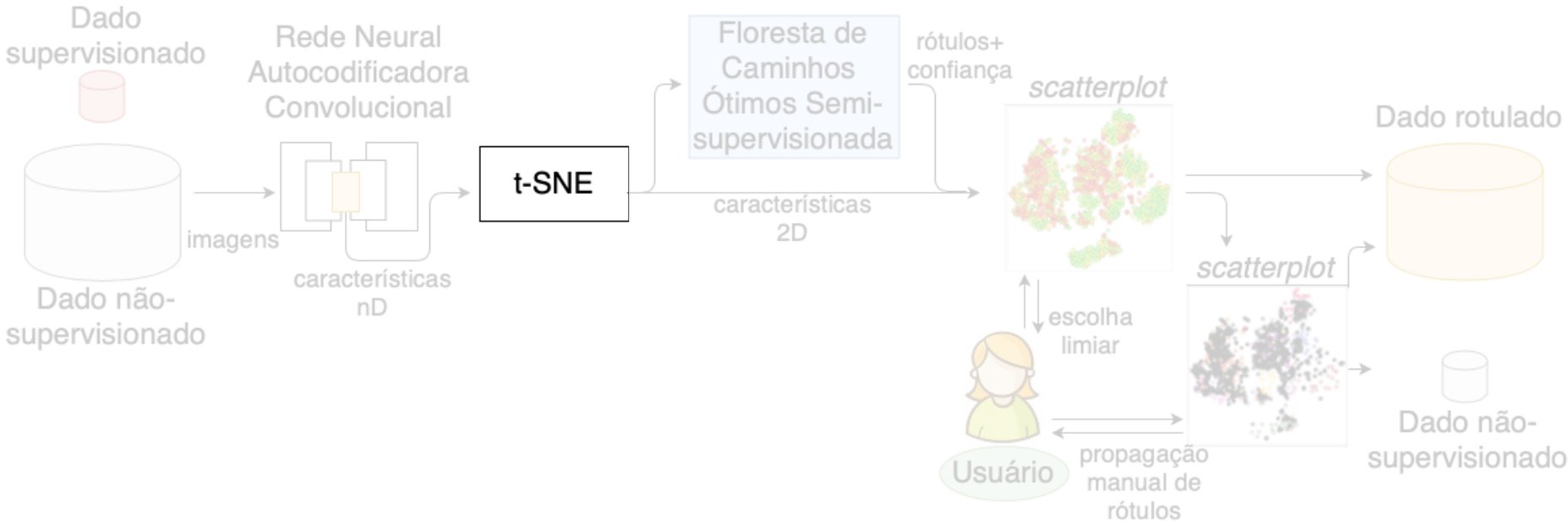


Fig. 1 − Método semi-automático proposto.

t-SNE

O t-SNE é um método de redução de dimensionalidade, apropriado para a visualização do espaço de características.

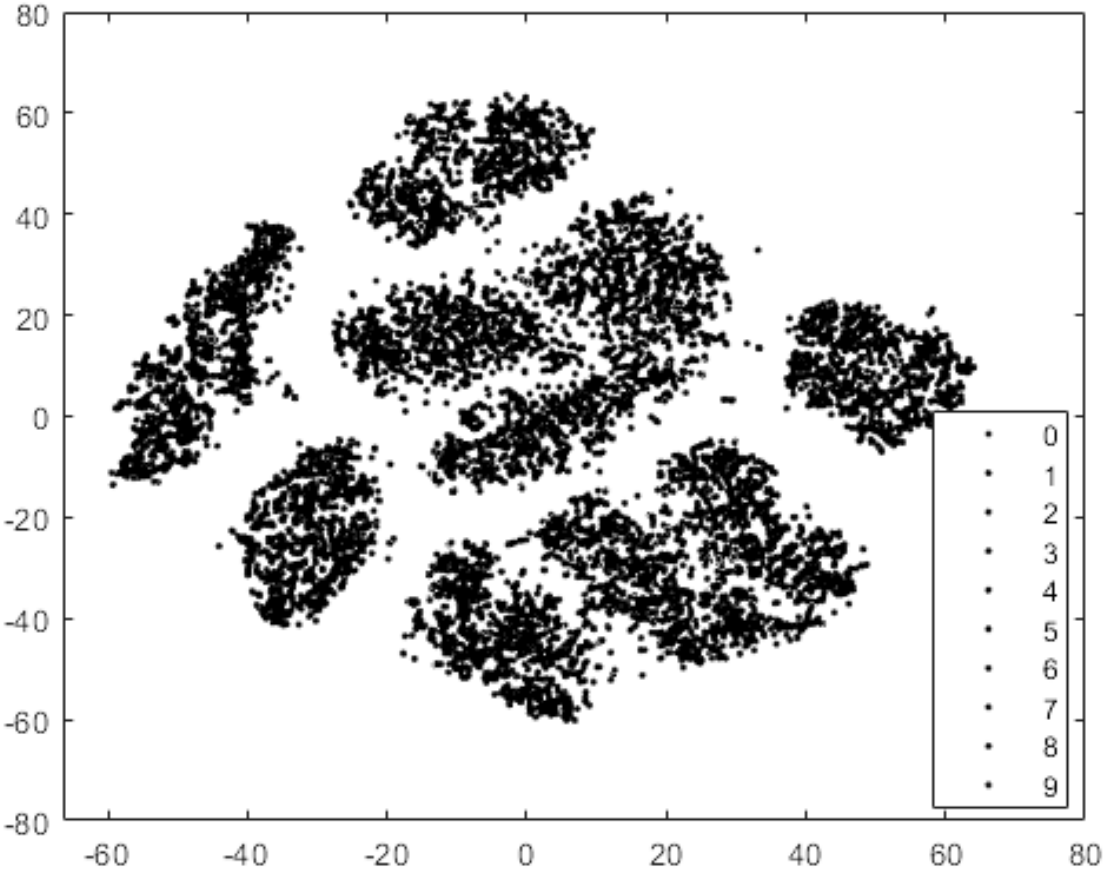


Figura reproduzida de: https://www.mathworks.com/help/stats/visualize-high-dimensional-data-using-t-sne.html

t-SNE

O t-SNE é um método de redução de dimensionalidade, apropriado para a visualização do espaço de características.

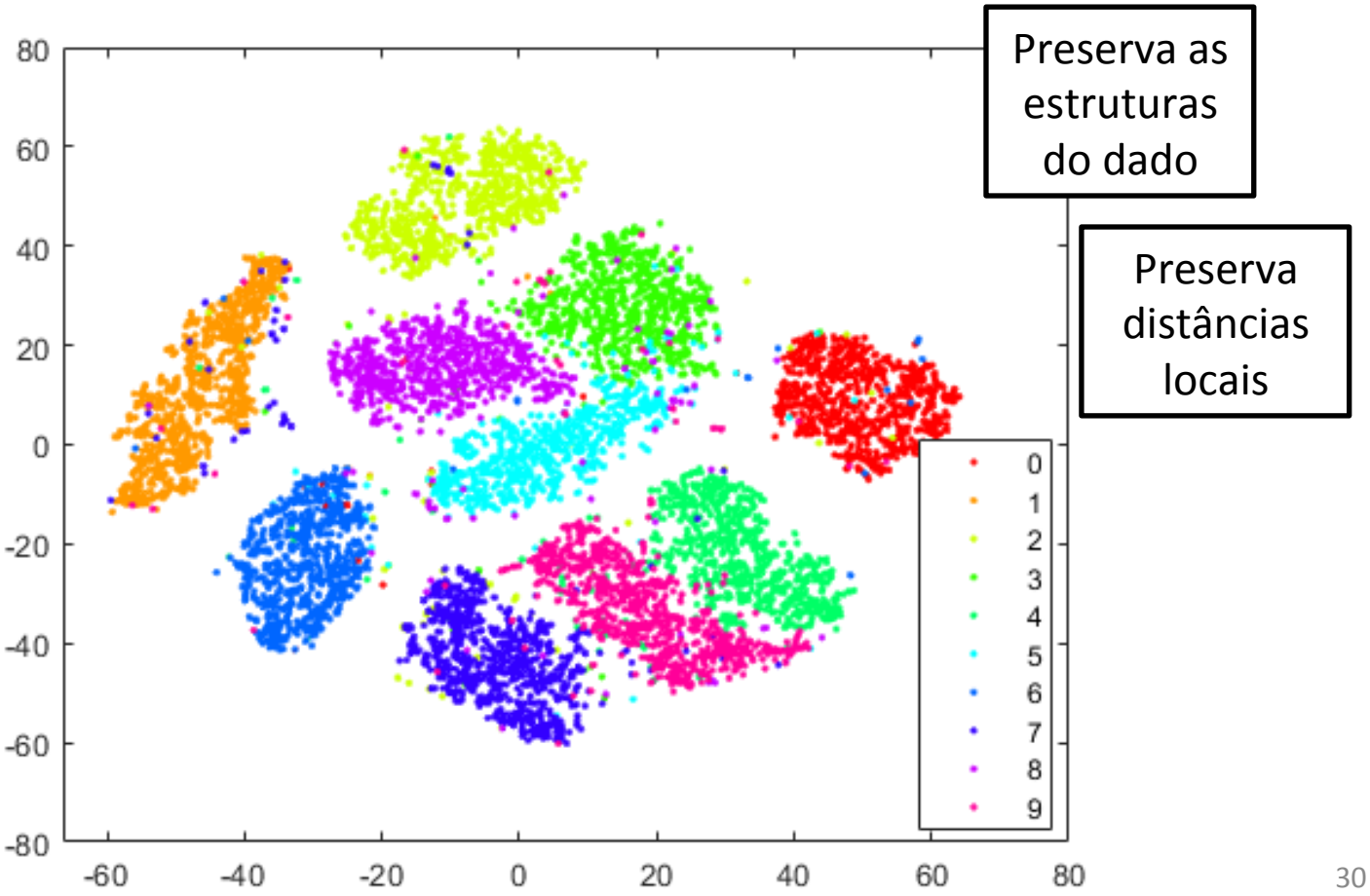


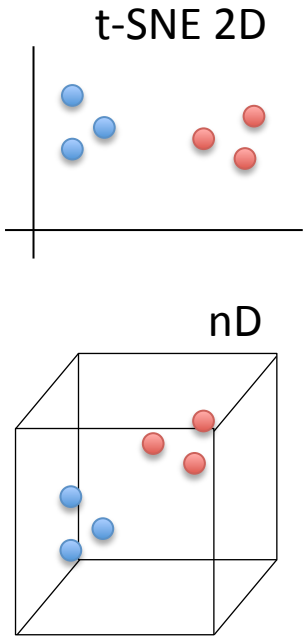
Figura reproduzida de: https://www.mathworks.com/help/stats/visualize-high-dimensional-data-using-t-sne.html

t-SNE

Rauber *et. al* mostraram como a projeção t-SNE se relaciona com classificadores de padrões.

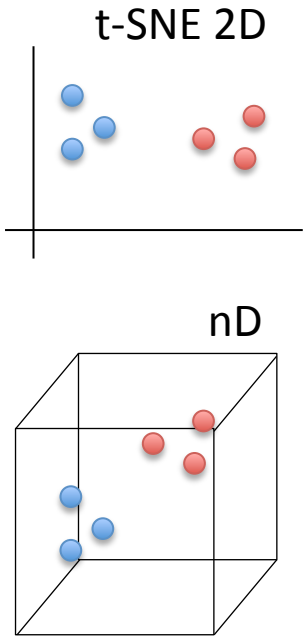
t-SNE

Rauber *et. al* mostraram como a projeção t-SNE se relaciona com classificadores de padrões.



t-SNE

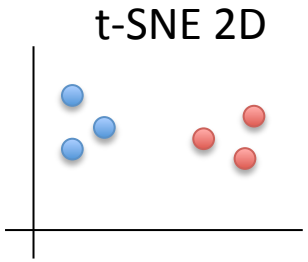
Rauber *et. al* mostraram como a projeção t-SNE se relaciona com classificadores de padrões.

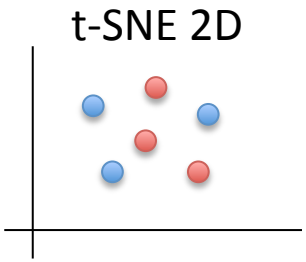


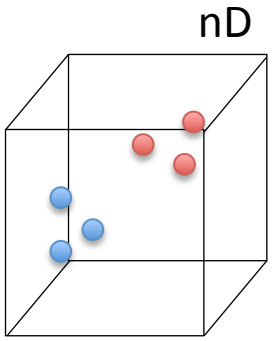
Altas acurácias
de classificação

t-SNE

Rauber *et. al* mostraram como a projeção t-SNE se relaciona com classificadores de padrões.



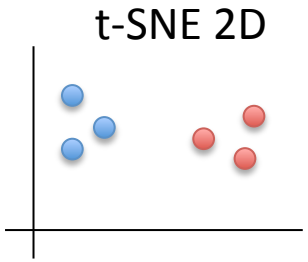


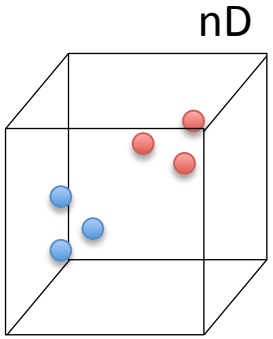


Altas acurácias
de classificação

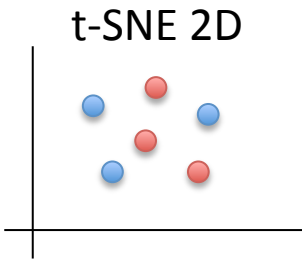
t-SNE

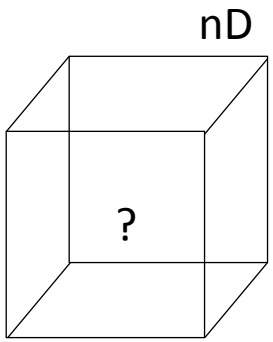
Rauber *et. al* mostraram como a projeção t-SNE se relaciona com classificadores de padrões.





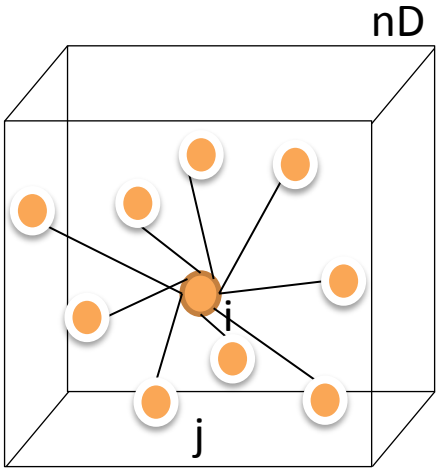
Altas acurácias de classificação



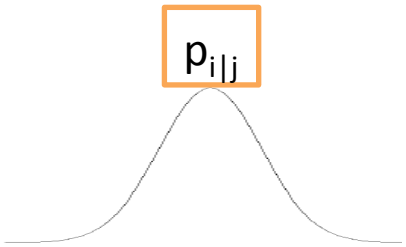


Não se pode afirmar

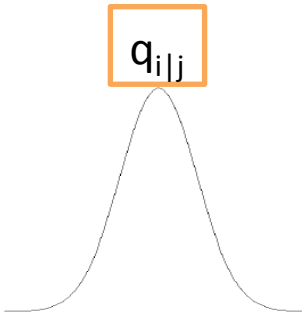
t-SNE



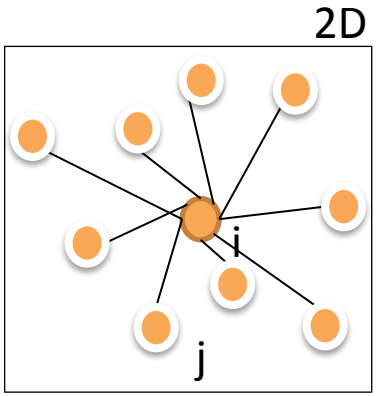
"n" vizinhos



Densidade
Probabilidade
Gaussiana

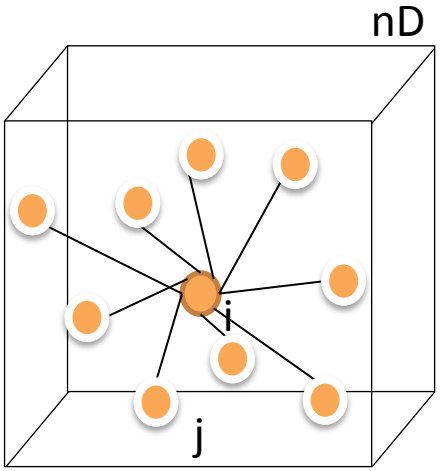


Densidade
Probabilidade
t-student



"n" vizinhos

t-SNE



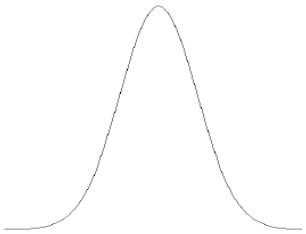
"n" vizinhos

Densidade
Probabilidade
Gaussiana

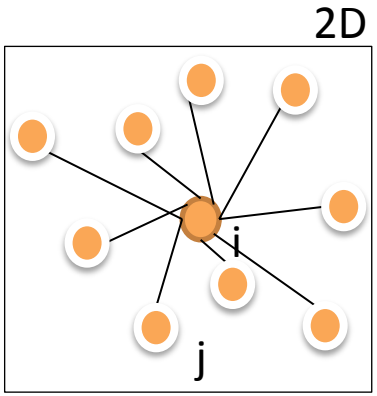
Aproximar as probabilidades

$p_{i|j}$ = $q_{i|j}$

Divergência de
Kullback Leiber

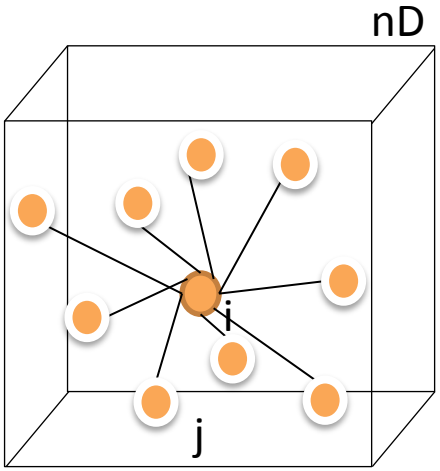


Densidade
Probabilidade
t-student



"n" vizinhos

t-SNE



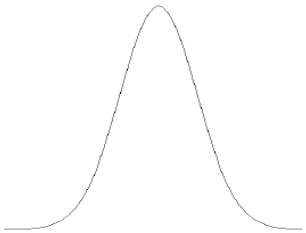
"n" vizinhos

Densidade
Probabilidade
Gaussiana

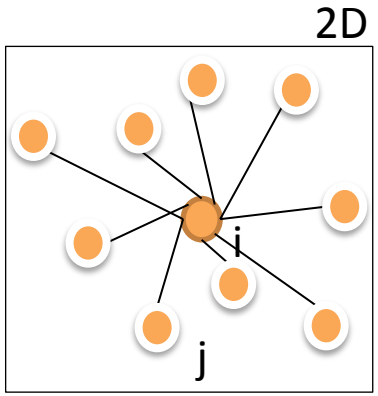
Aproximar as probabilidades

$p_{i|j}$ = $q_{i|j}$

Divergência de
Kullback Leiber

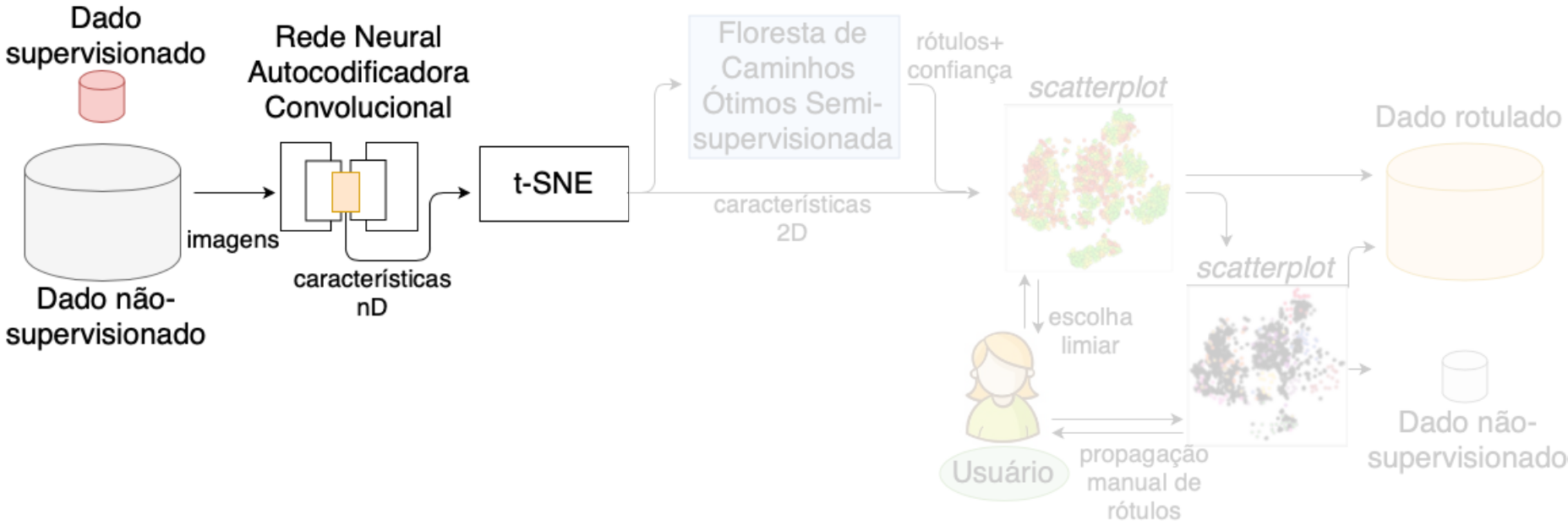


Densidade
Probabilidade
t-student

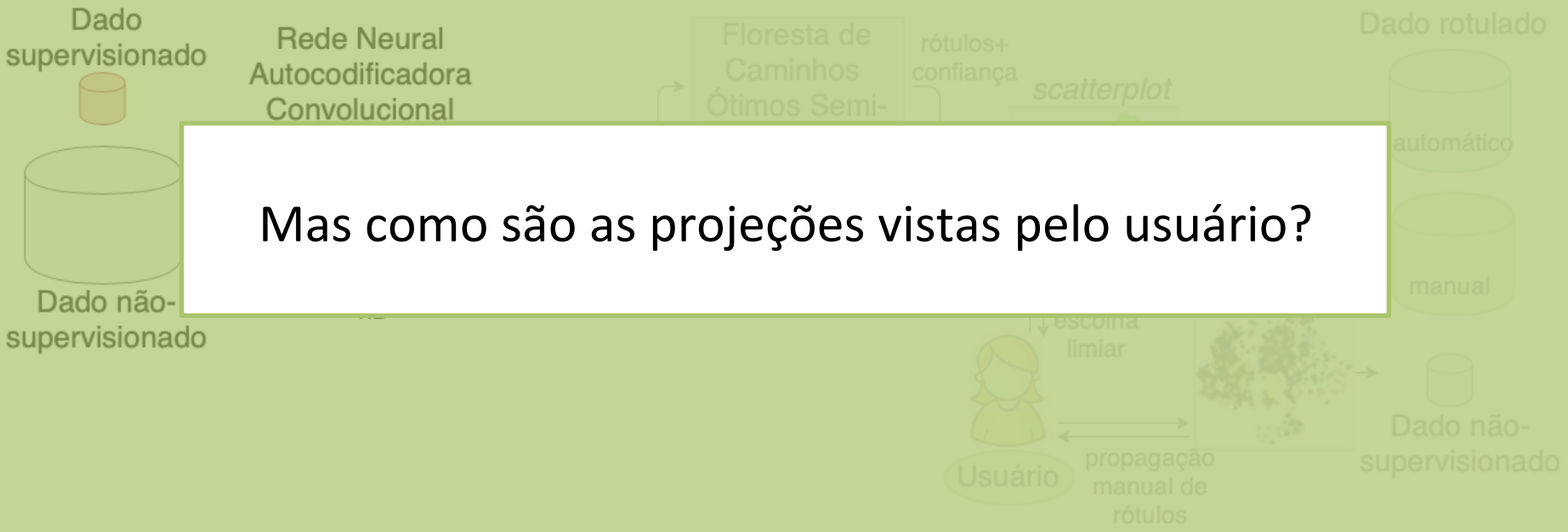


"n" vizinhos

Propagação Semi-automática de rótulos (SALP)



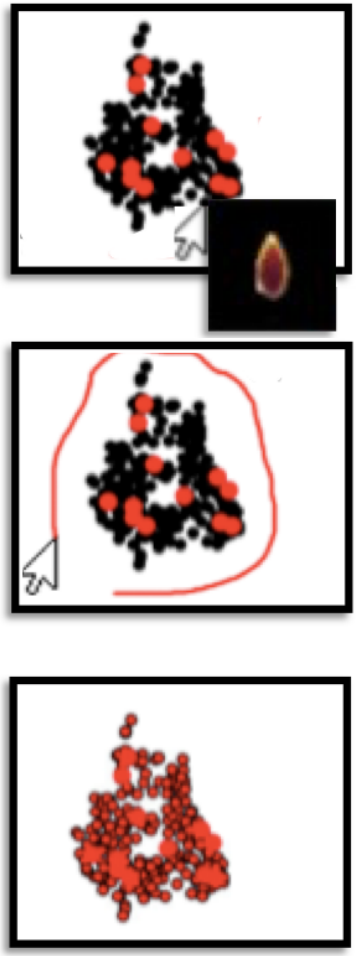
Propagação Semi-automática de rótulos (SALP)



Mas como são as projeções vistas pelo usuário?

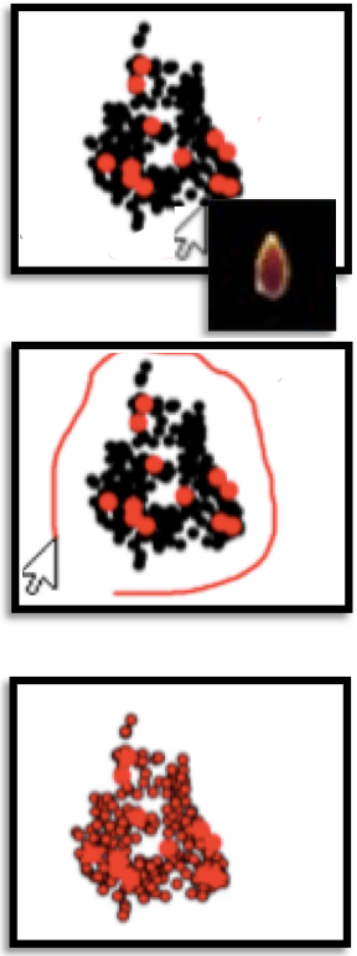
Intuição da propagação manual de rótulos (ILP)

Ferramentas de Propagação manual

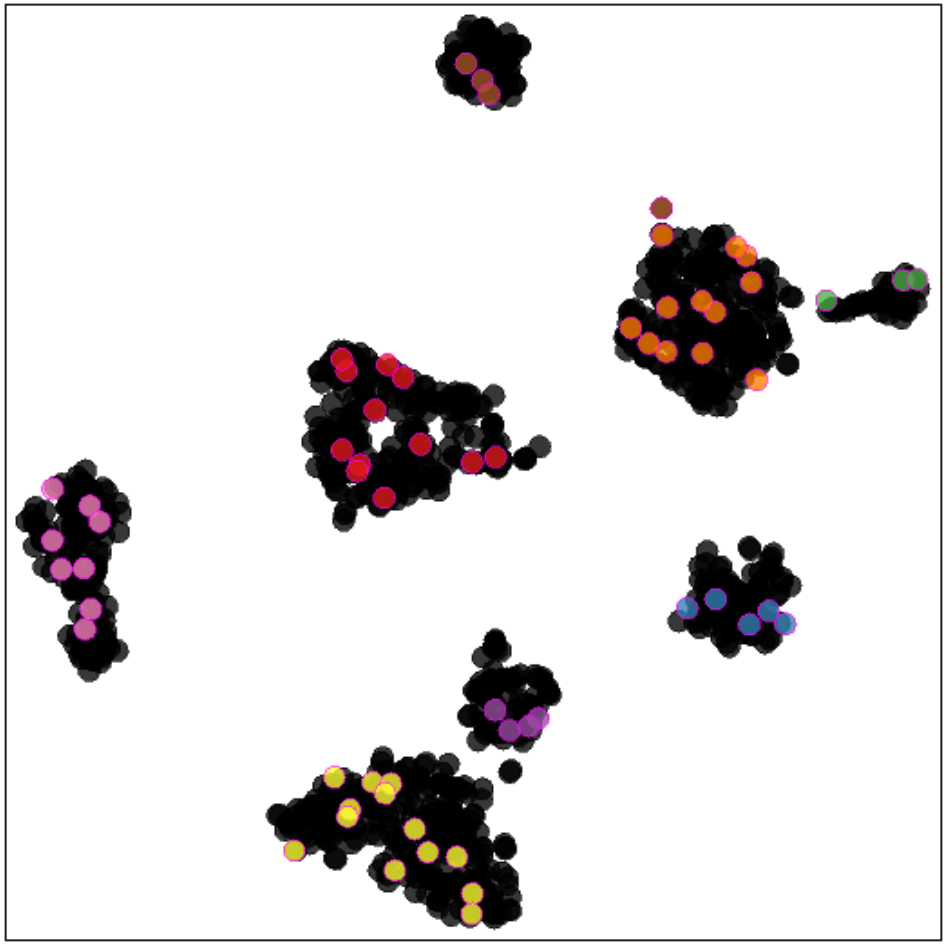


Intuição da propagação manual de rótulos (ILP)

Ferramentas de Propagação manual

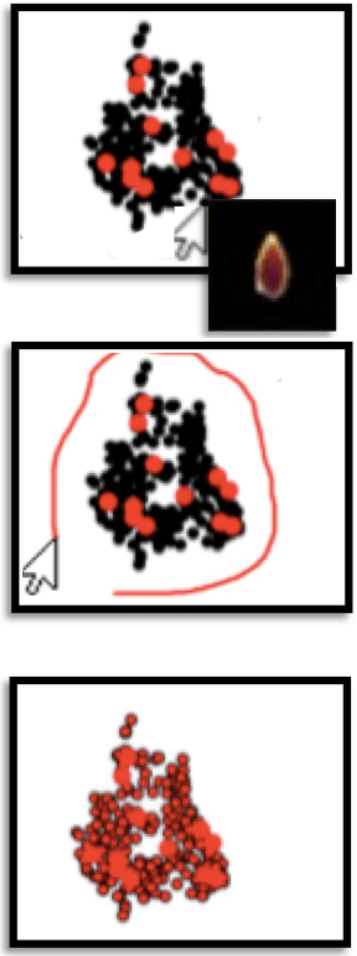


Projeção t-SNE 2D

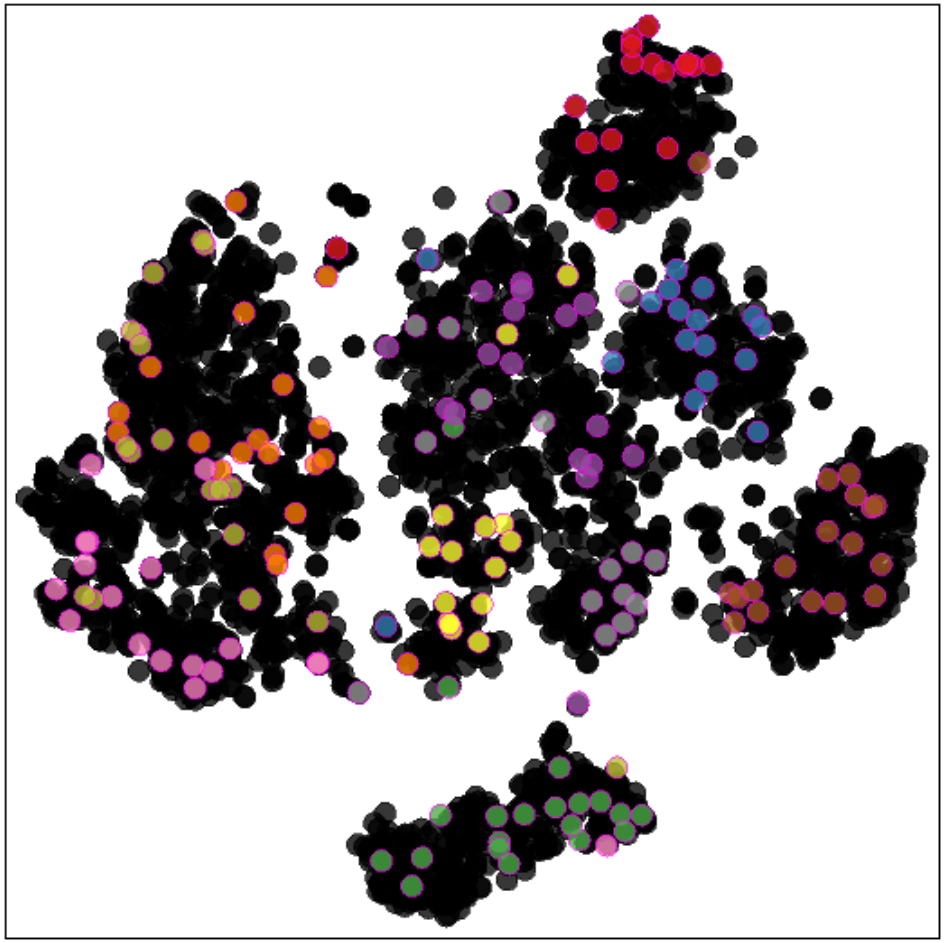


Intuição da propagação manual de rótulos (ILP)

Ferramentas de Propagação manual

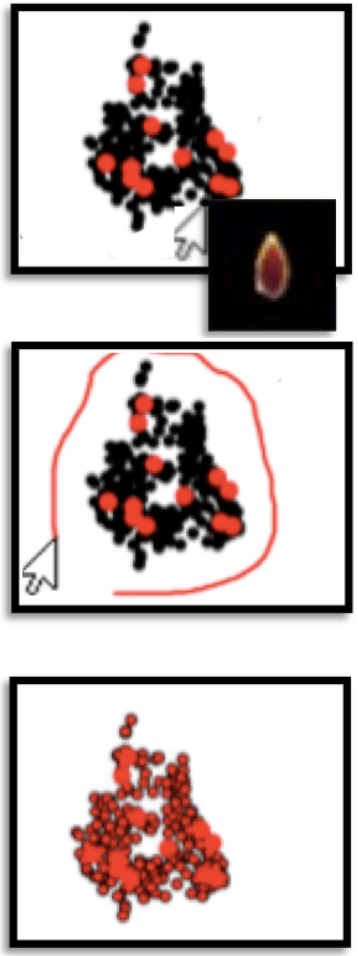


Projeção t-SNE 2D

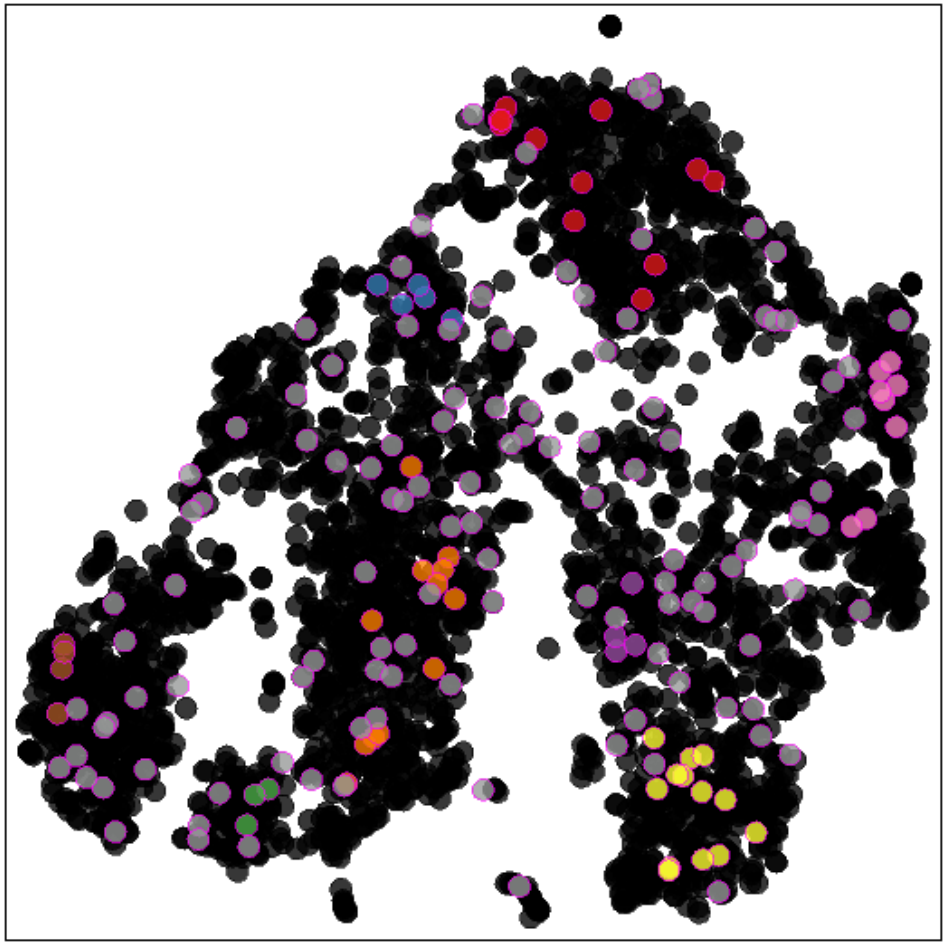


Intuição da propagação manual de rótulos (ILP)

Ferramentas de Propagação manual

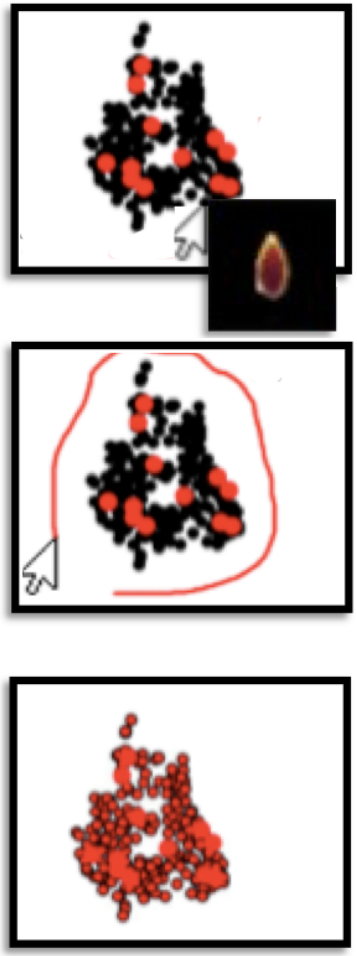


Projeção t-SNE 2D

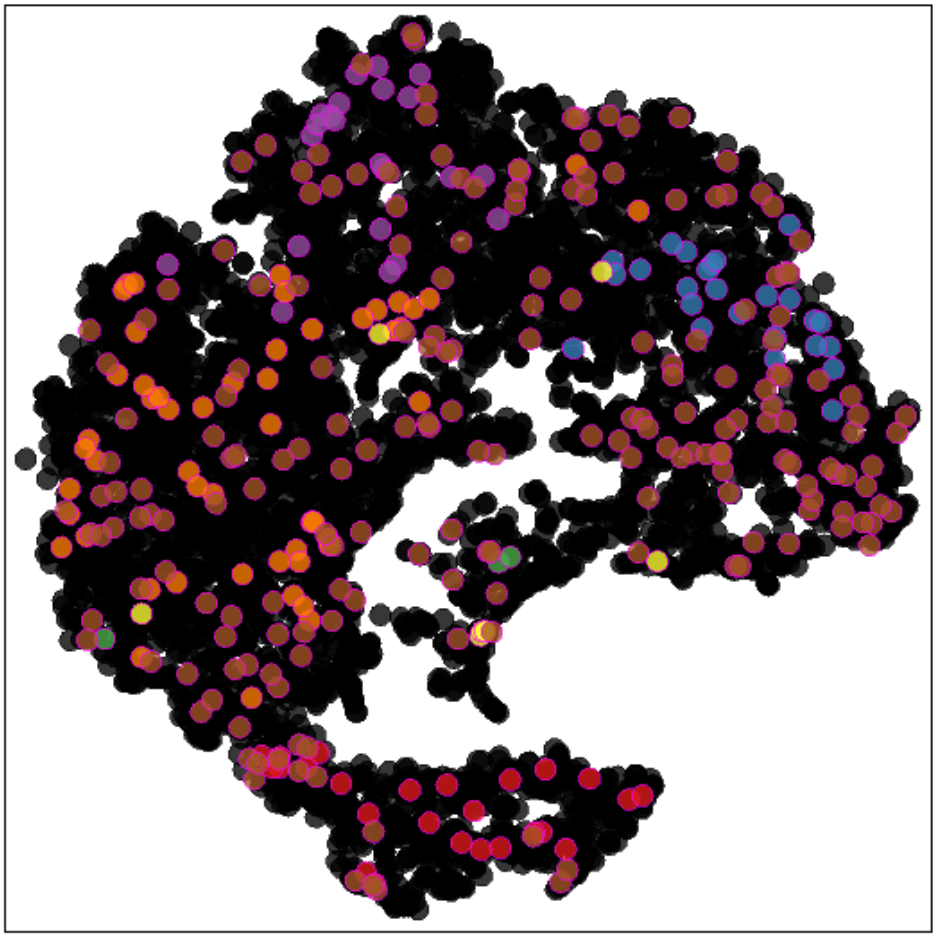


Intuição da propagação manual de rótulos (ILP)

Ferramentas de Propagação manual



Projeção t-SNE 2D



Intuição da Propagação Manual

Ferramentas de Propagação manual

Projeção t-SNE 2D

Como reduzir o esforço do usuário?

Propagação Semi-automática de rótulos (SALP)

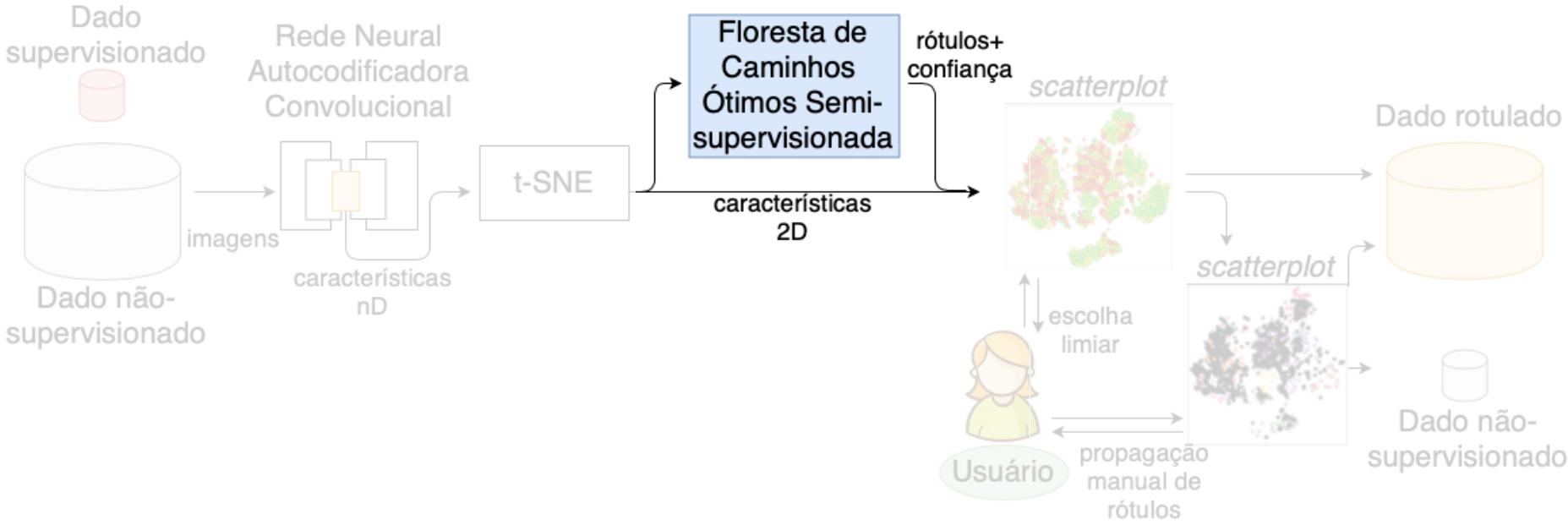


Fig. 1 − Método semi-automático proposto.

Estimação de rótulos semi-supervisionada

Algoritmo de classificação semi-supervisionado Floresta de Caminhos Ótimos (OPF-Semi) é utilizado para obter a informação de **confiança**.

- Amostras mais fáceis de serem classificadas.
- Amostras mais difíceis de serem classificadas.

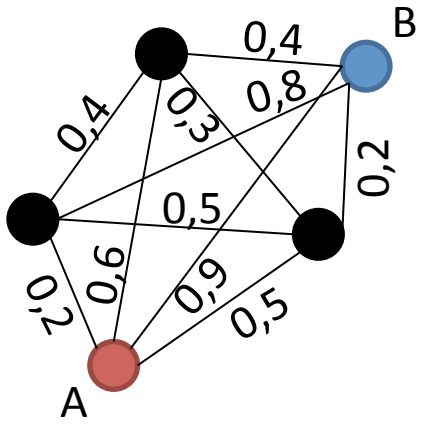
Estimação de rótulos semi-supervisionada

Algoritmo de classificação semi-supervisionado Floresta de Caminhos Ótimos (OPF-Semi) é utilizado para obter a informação de **confiança**.

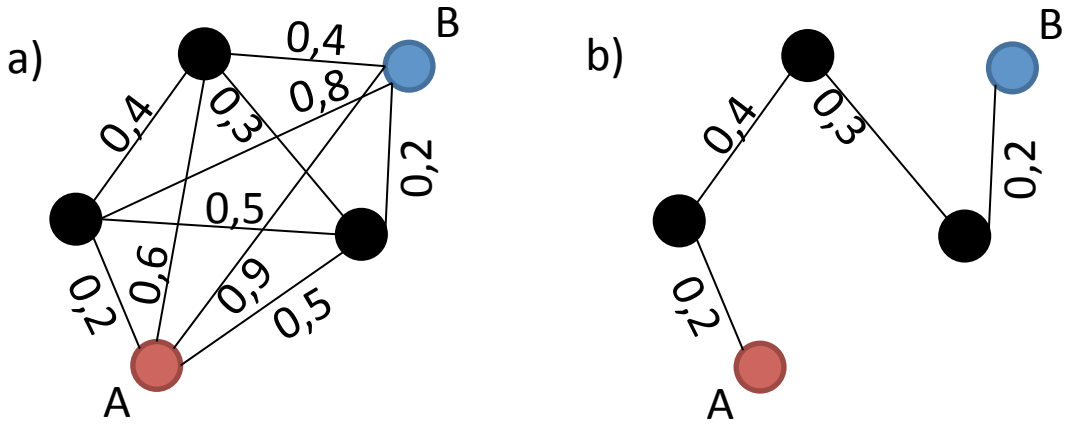
- Amostras mais fáceis de serem classificadas. → Classificador
- Amostras mais difíceis de serem classificadas. → Usuário

OPF-Semi

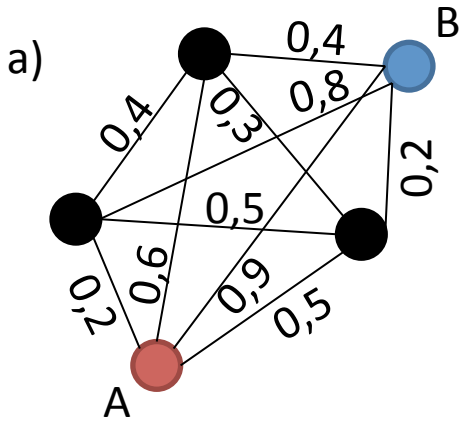
a)

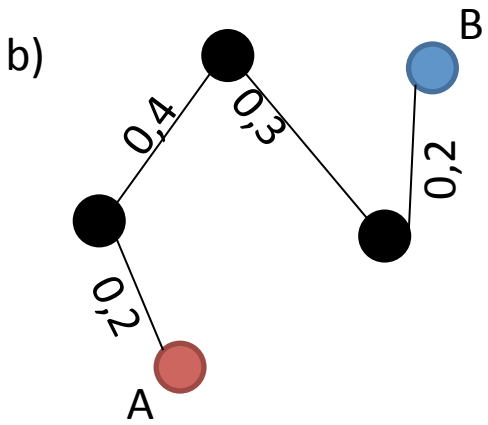


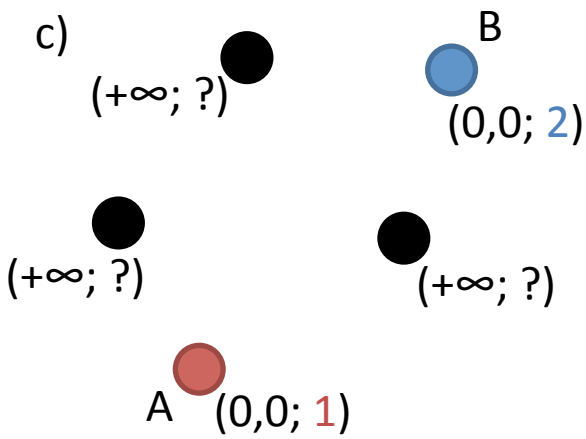
OPF-Semi



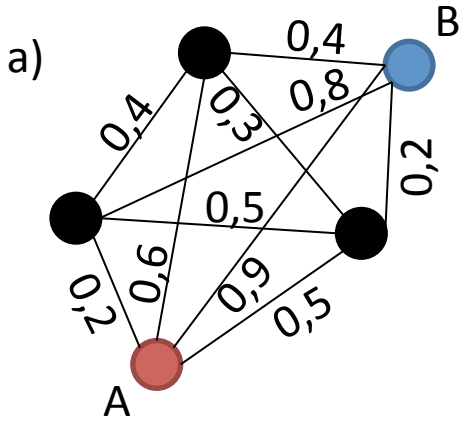
OPF-Semi

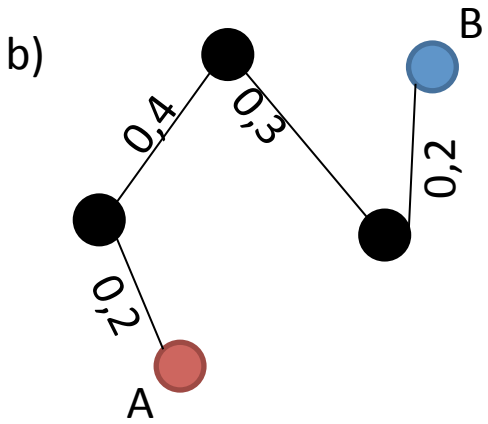


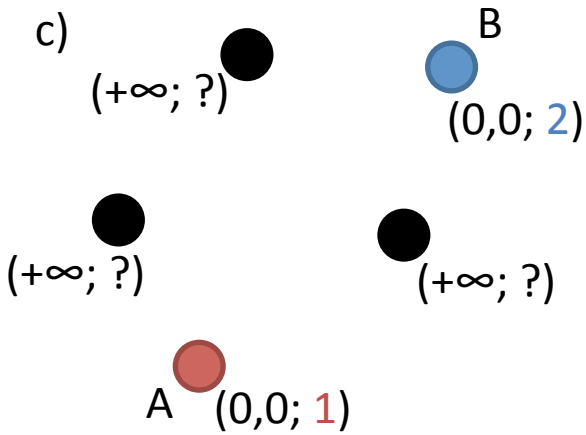


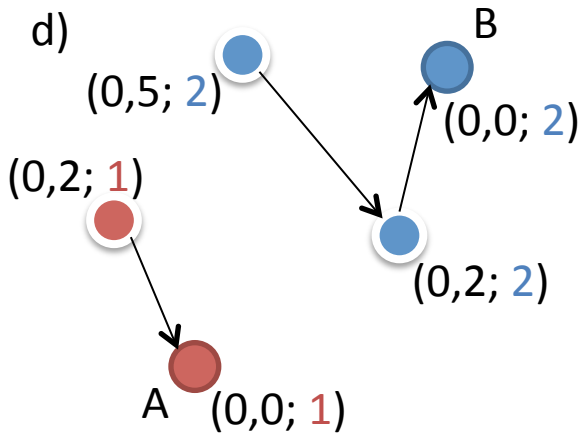


OPF-Semi



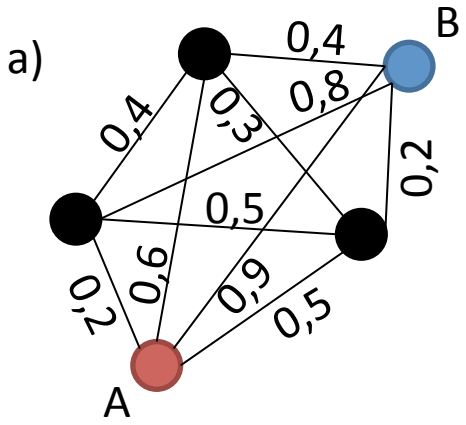


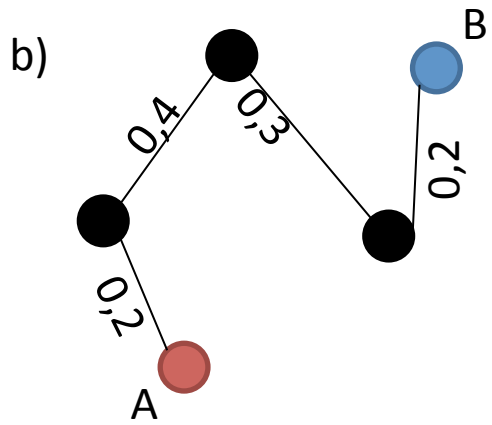


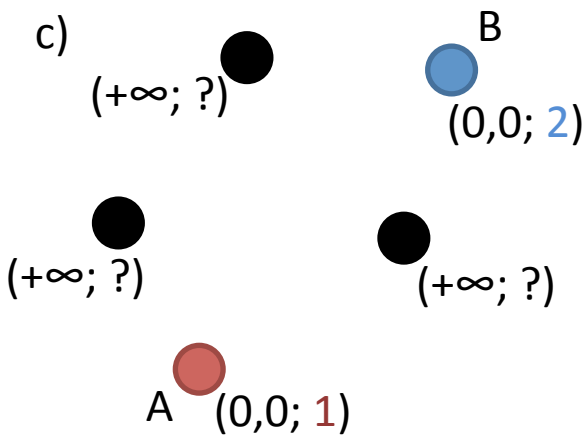


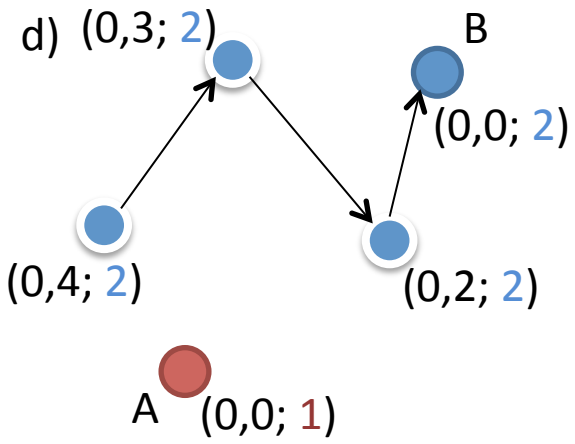
Uma floresta com competição é criada.

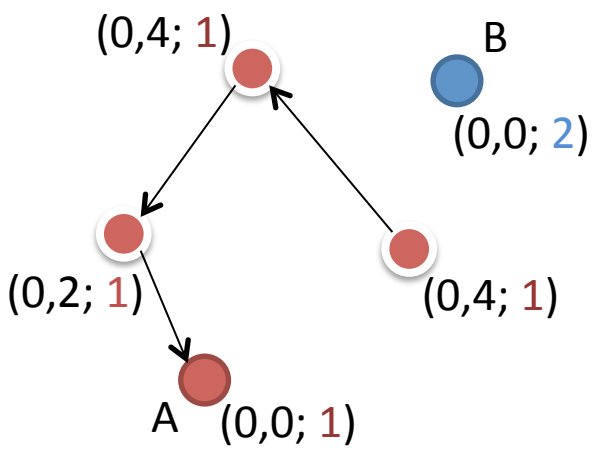
Confiança pelo OPF-Semi









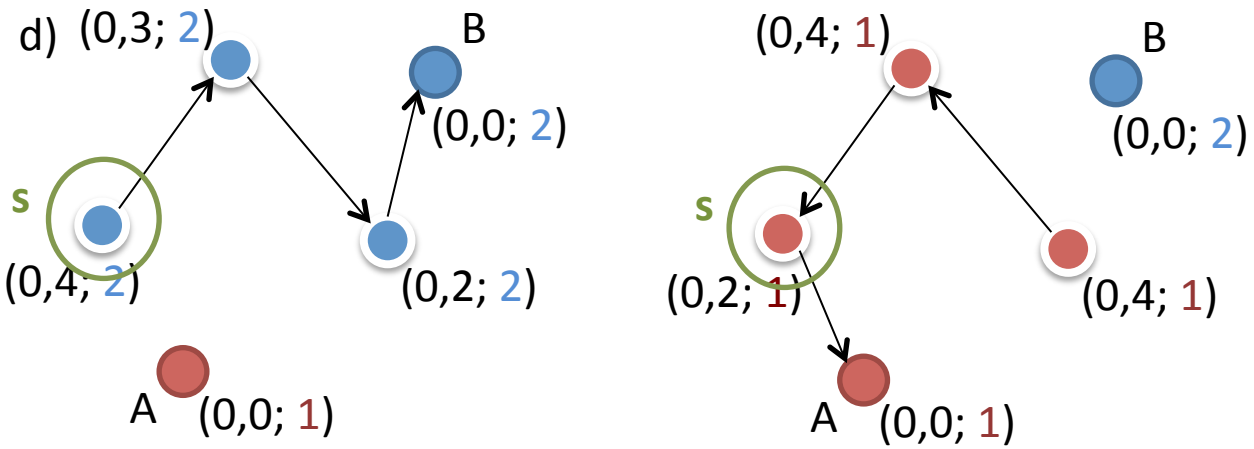


"L" florestas são criadas.

Confiança pelo OPF-Semi

A **confiança** da amostra "**s**" é dada pelo segundo melhor custo oferecido.

Confiança = 0,4 / (0,2 + 0,4)
= **0,67**



Estimação de rótulos semi-supervisionada

Depois de obtida a confiança, é necessário **definir** as amostras que são mais fáceis e mais difíceis de serem classificadas.

Limiar de confiança

Propagação Semi-automática de rótulos (SALP)

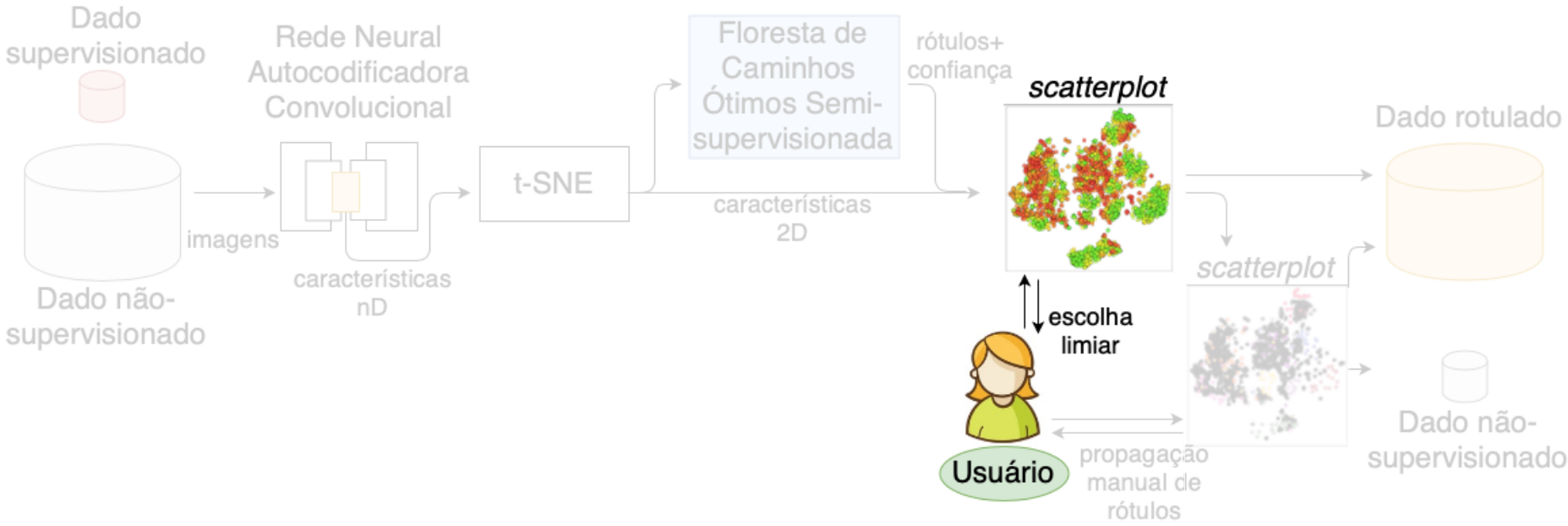
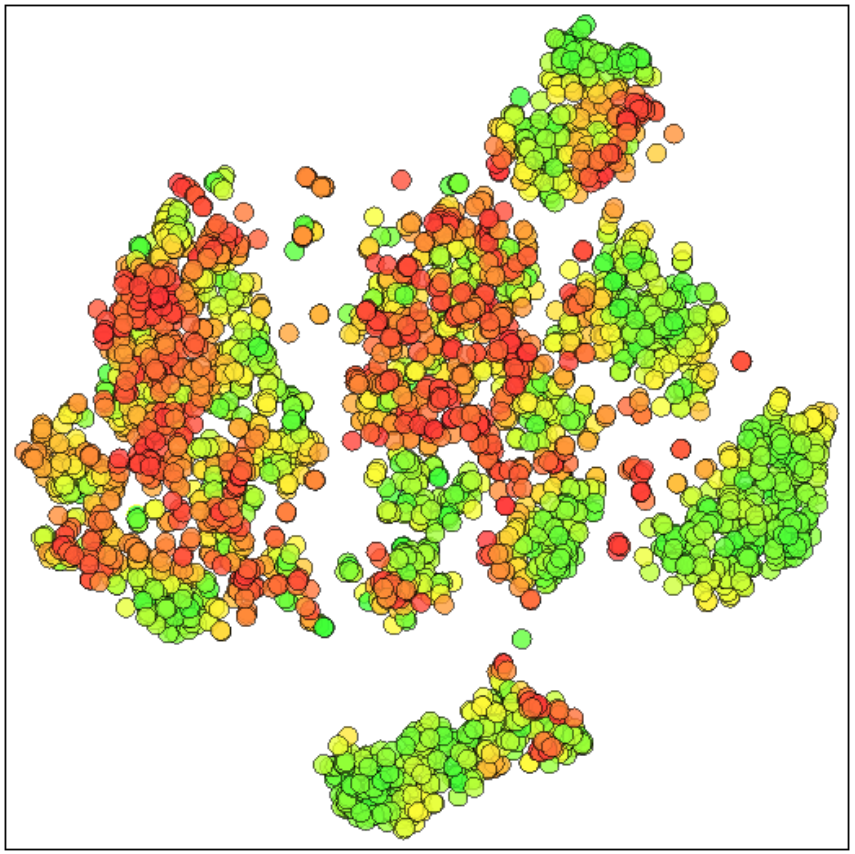


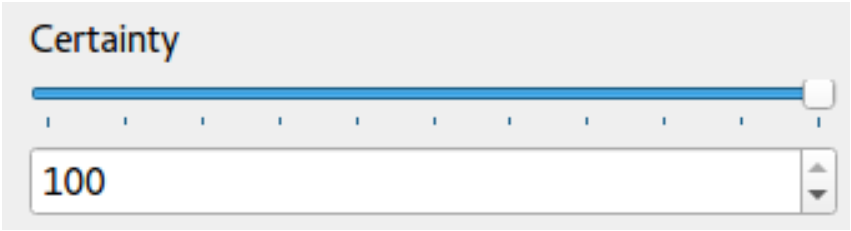
Fig. 1 − Método semi-automático proposto.

Limiar de Confiança



Limiar de Confiança

Usuário escolhe um limiar de confiança.



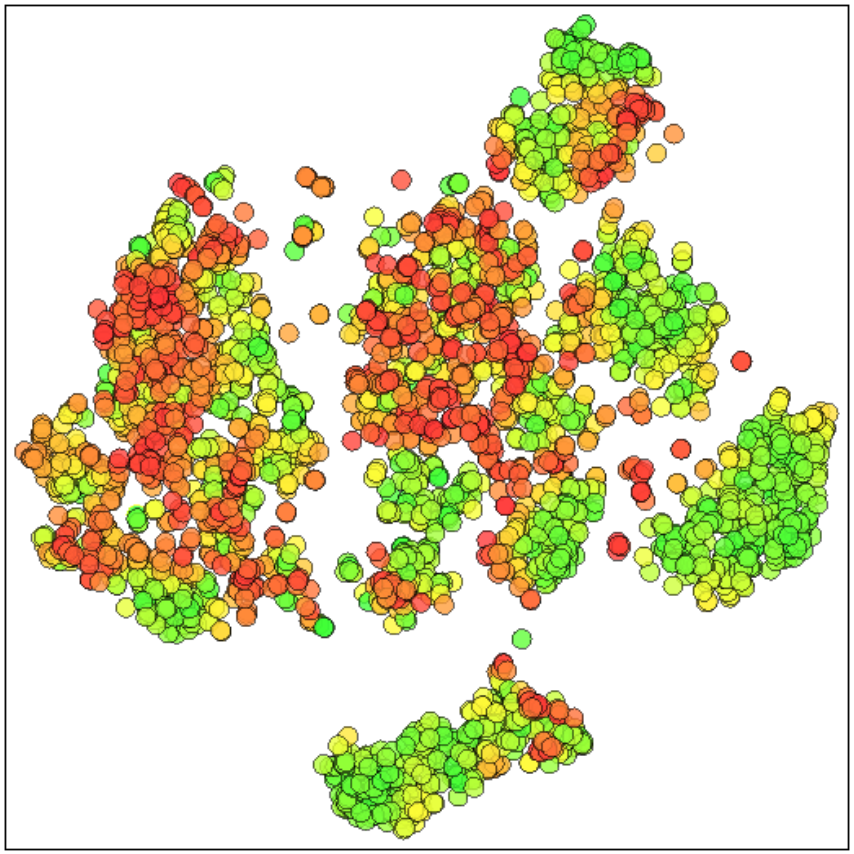
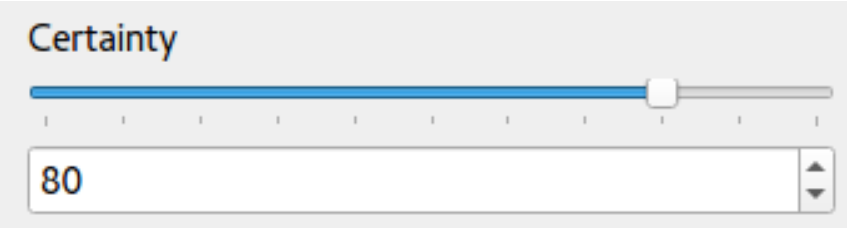


Fig. 3 − Ferramenta de deslizar que permite escolher o limiar de confiança.
E projeção t-SNE do espaço de características, colorida pela confiança do OPF-Semi.

Limiar de Confiança

Usuário escolhe um limiar de confiança.



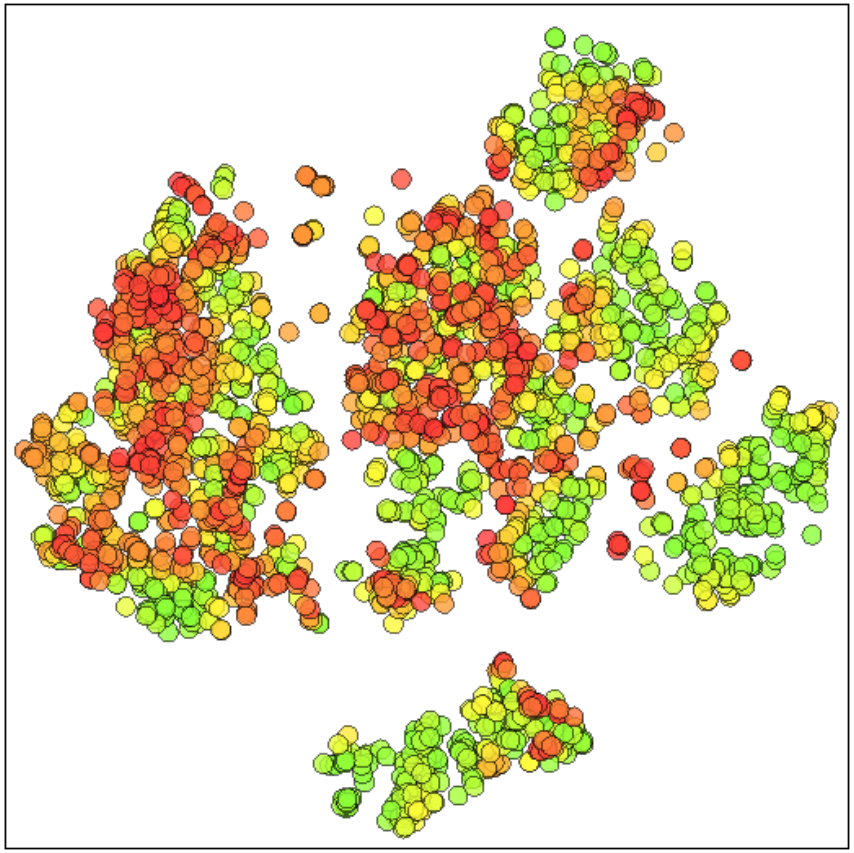
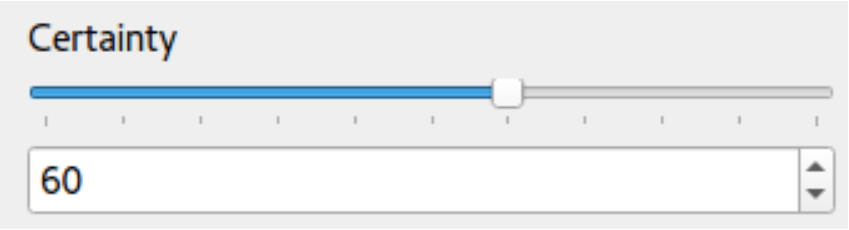


Fig. 3 − Ferramenta de deslizar que permite escolher o limiar de confiança.
E projeção t-SNE do espaço de características, colorida pela confiança do OPF-Semi.

Limiar de Confiança

Usuário escolhe um limiar de confiança.



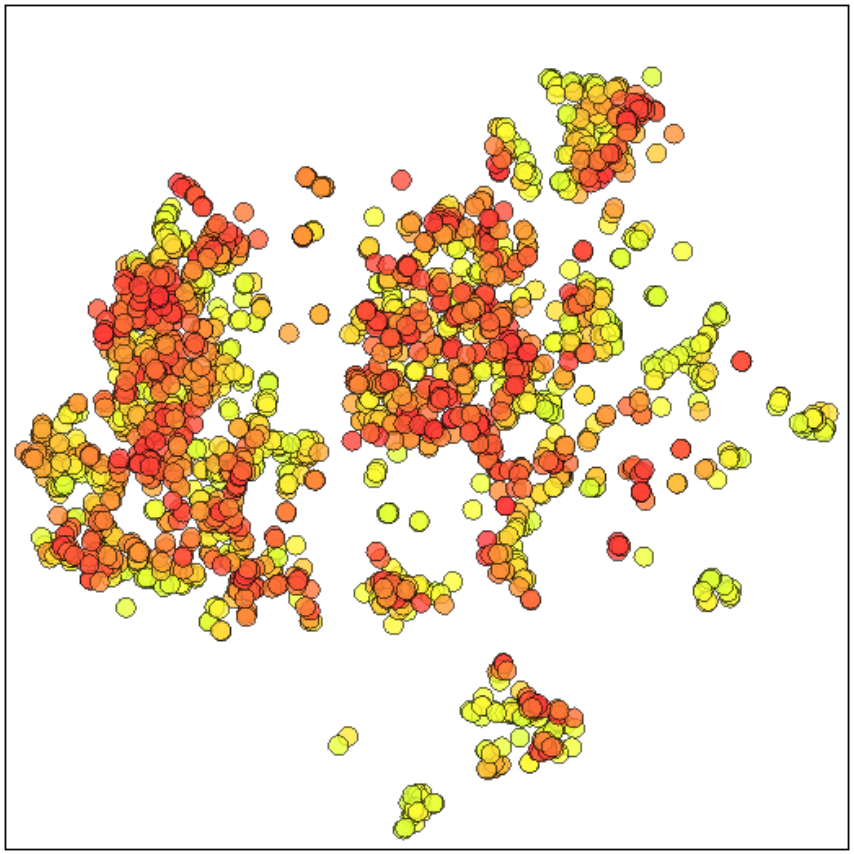
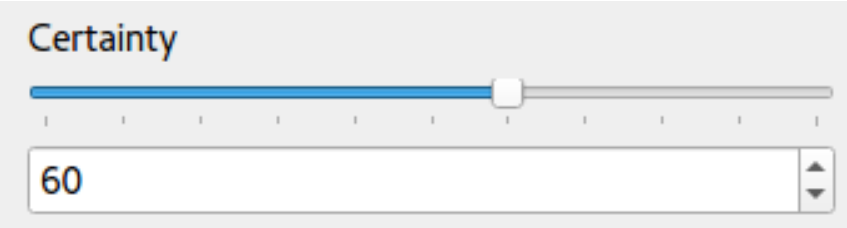


Fig. 3 − Ferramenta de deslizar que permite escolher o limiar de confiança.
E projeção t-SNE do espaço de características, colorida pela confiança do OPF-Semi.

Limiar de Confiança

Usuário escolhe um limiar de confiança.



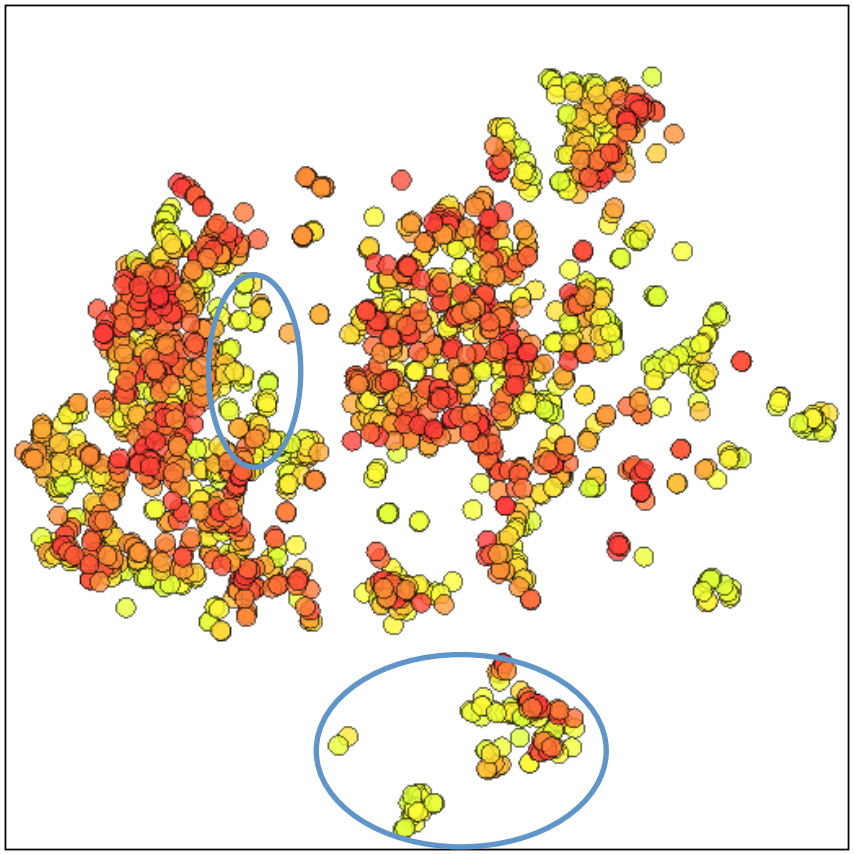


Fig. 3 − Ferramenta de deslizar que permite escolher o limiar de confiança.
E projeção t-SNE do espaço de características, colorida pela confiança do OPF-Semi.

Propagação Semi-automática de rótulos (SALP)

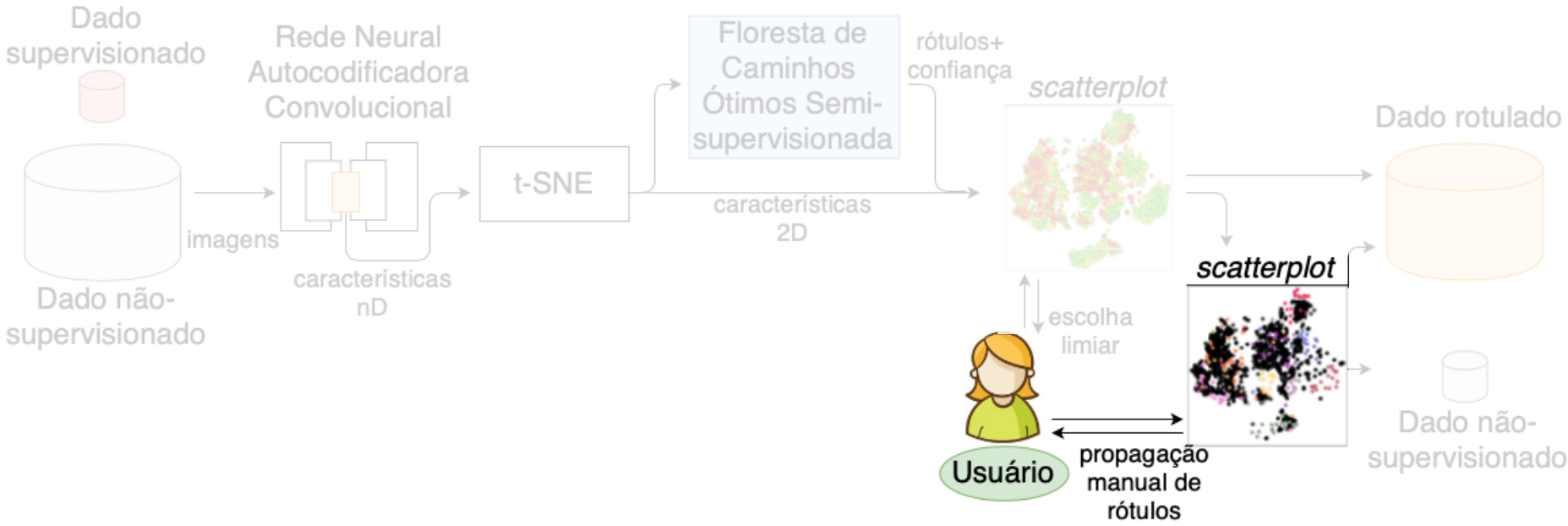
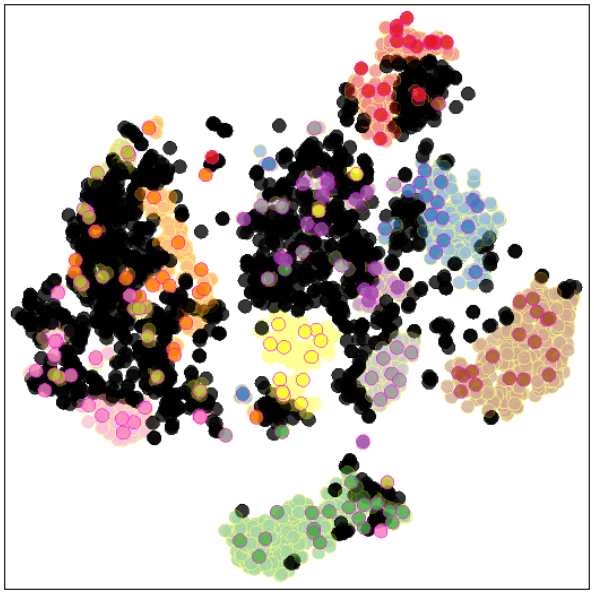


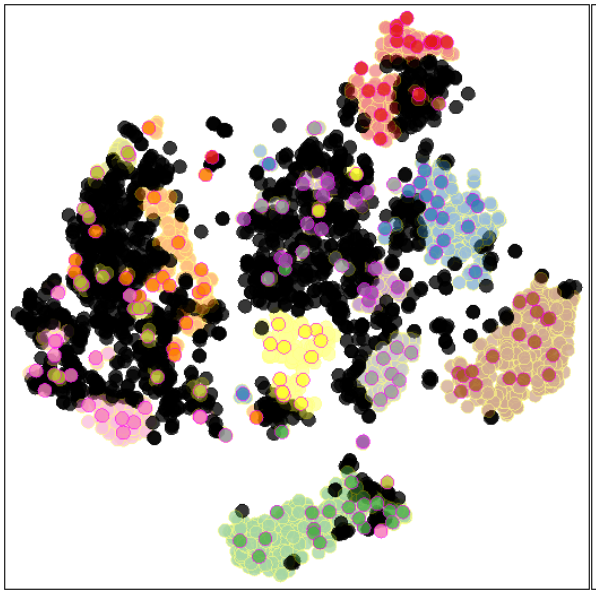
Fig. 1 − Método semi-automático proposto.

Propagação de rótulos manual (ILP)

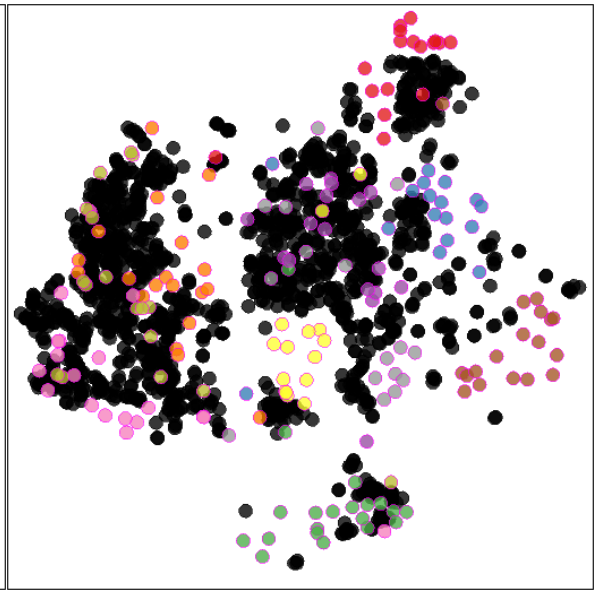


Amostras mais fáceis rotuladas pelo OPF-Semi.

Propagação de rótulos manual (ILP)

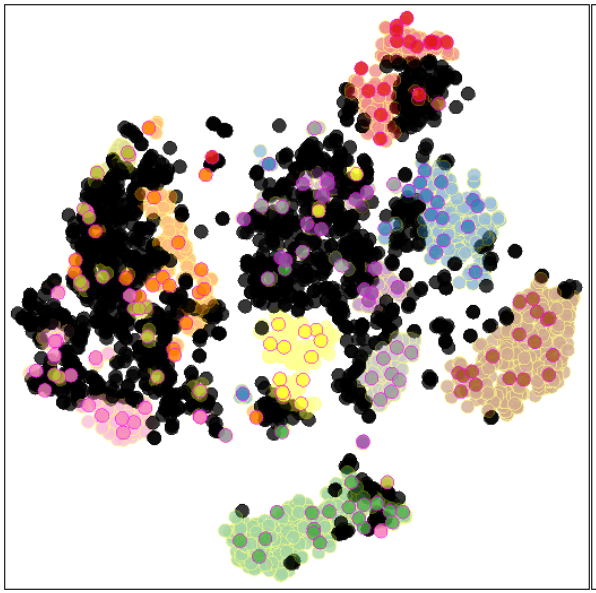


Amostras mais fáceis rotuladas pelo OPF-Semi.

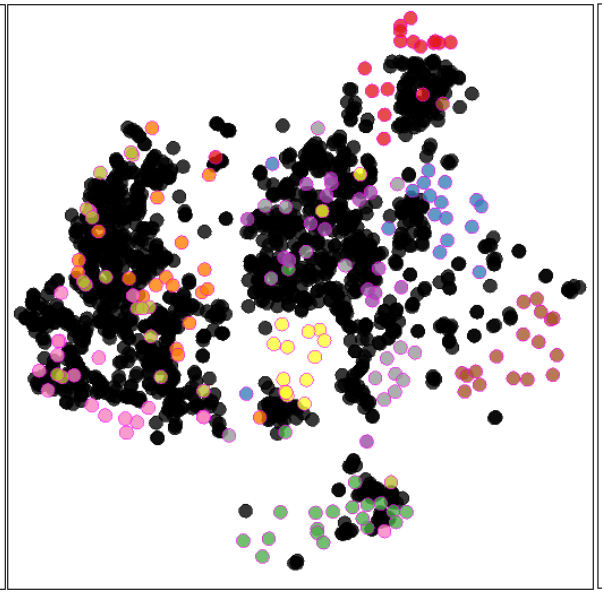


Amostras mais difíceis para o usuário anotar.

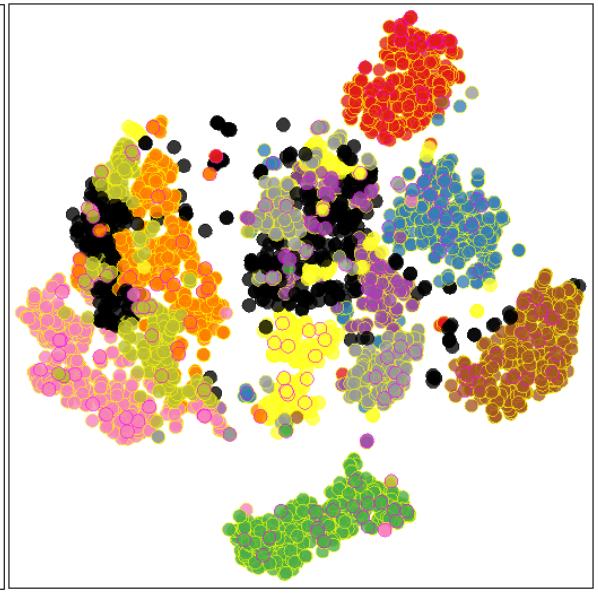
Propagação de rótulos manual (ILP)



Amostras mais fáceis rotuladas pelo OPF-Semi.



Amostras mais difíceis para o usuário anotar.



Amostras rotuladas pelo OPF-Semi e usuário.

Principais questões

- Qual é o melhor espaço para a propagação de rótulos automática?
- Qual limiar de confiança escolher?
- Qual abordagem (automática, manual, semi-automática) melhor propaga rótulos?
- Qual o valor agregado pelo método semi-automático proposto?
- Como os resultados dependem da qualidade da projeção?

Bases de Dados

- *MNIST:*
  - 5000 imagens, 10 classes
  - Dimensão: 28x28 (784)
- *Parasitos (in-house):*
    - Helminth larvae: 3514 imagens, 1 classe + impureza
    - Helminth eggs: 5112 imagens, 8 classes + impureza
    - Protozoan cysts: 9568 imagens, 6 classes + impureza
  - Com e sem impureza: 5 bases
  - Dimensão: 3x200x200 (120.000)

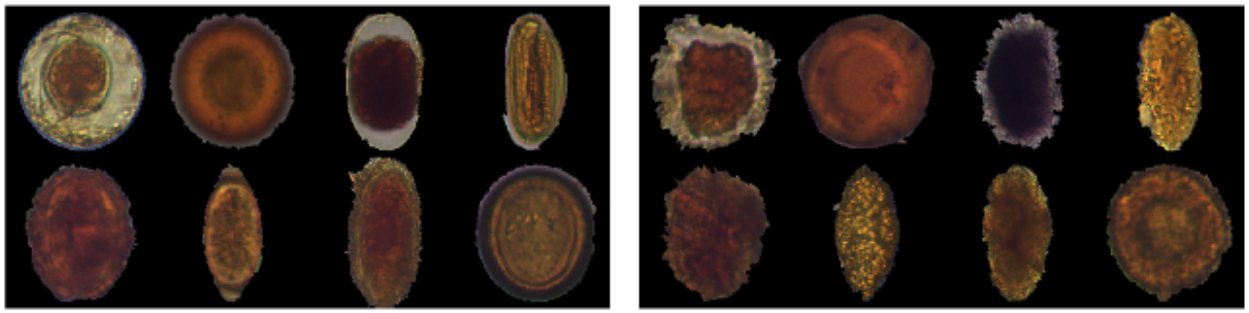


Fig. 5 − Base de dados de Parasitos: classes de Parasitos e impurezas.

Configuração Experimental

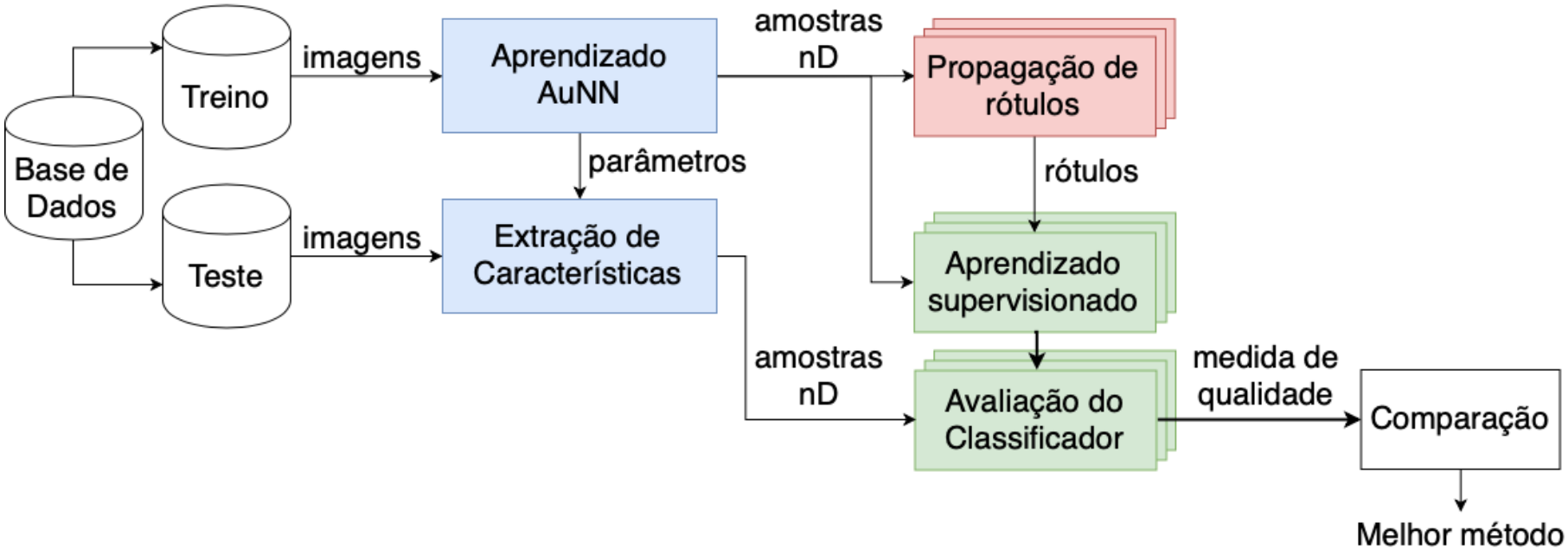


Fig. 6 − Validação do método semi-automático proposto.

Configuração Experimental

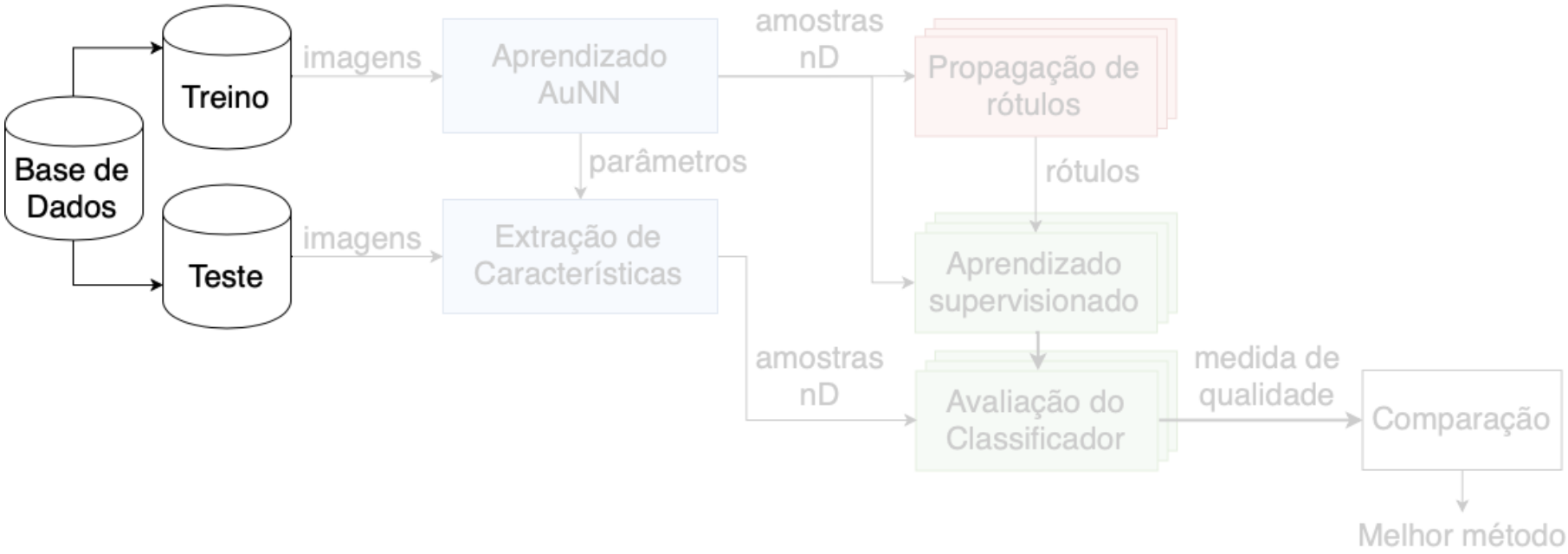


Fig. 6 − Validação do método semi-automático proposto.

Configuração Experimental

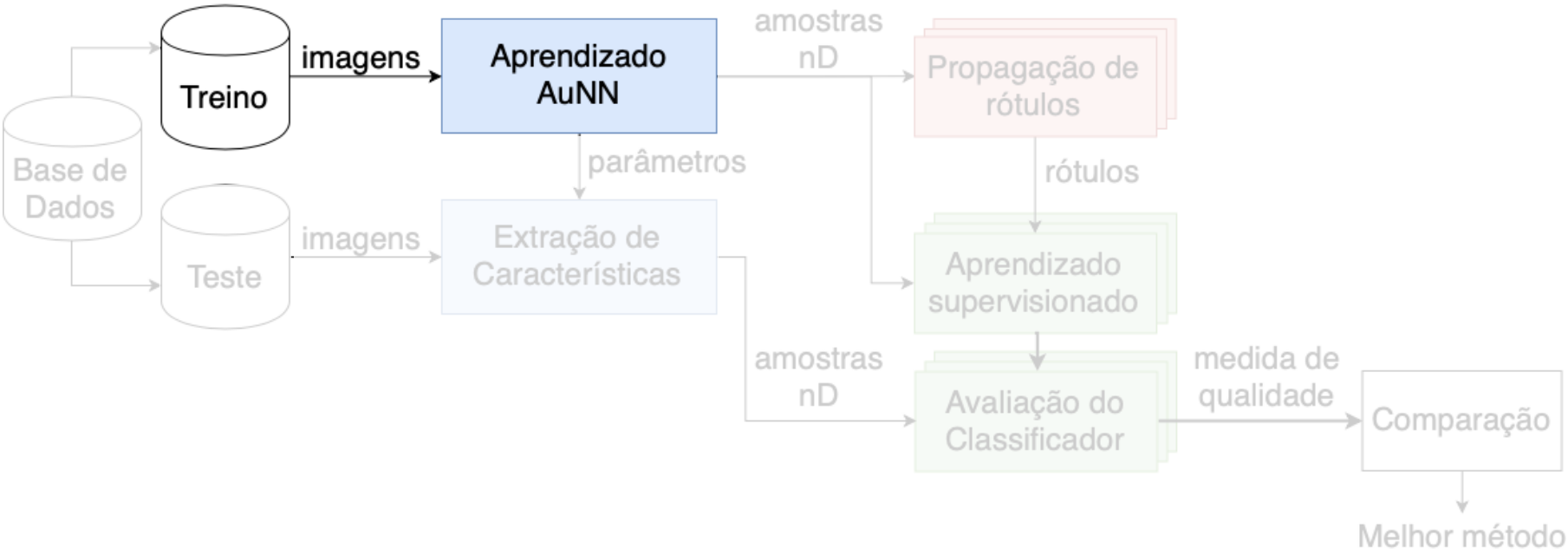


Fig. 6 − Validação do método semi-automático proposto.

Configuração Experimental

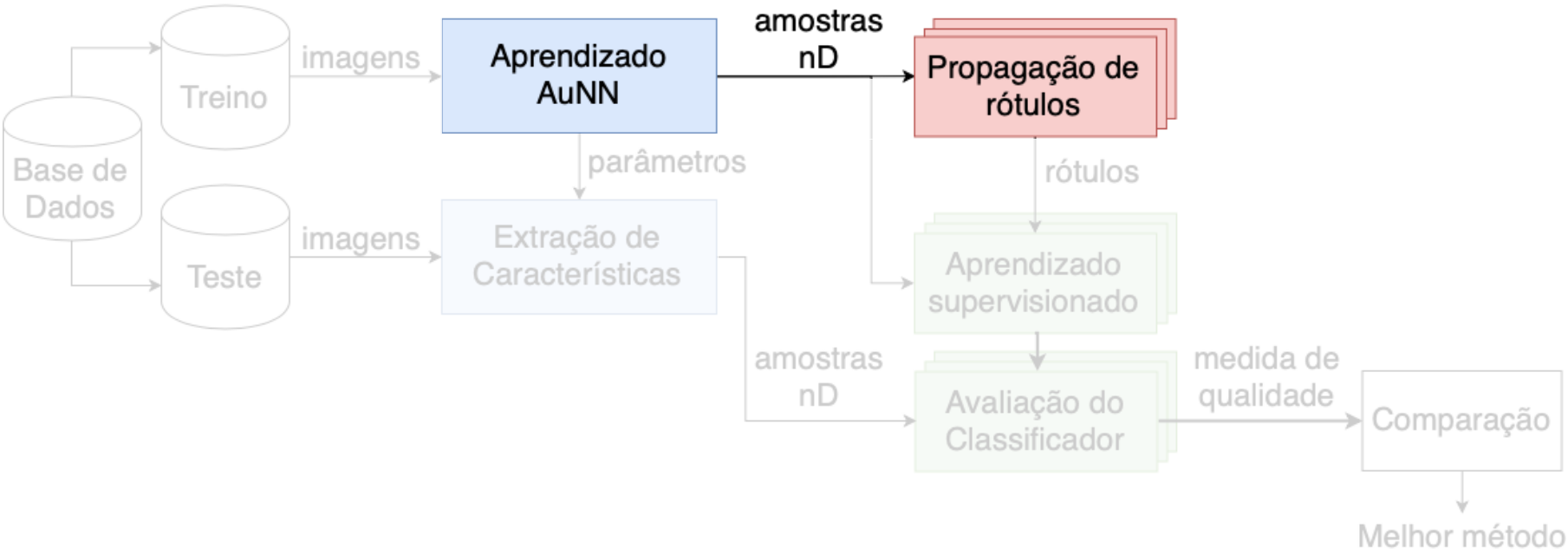


Fig. 6 − Validação do método semi-automático proposto.

Configuração Experimental

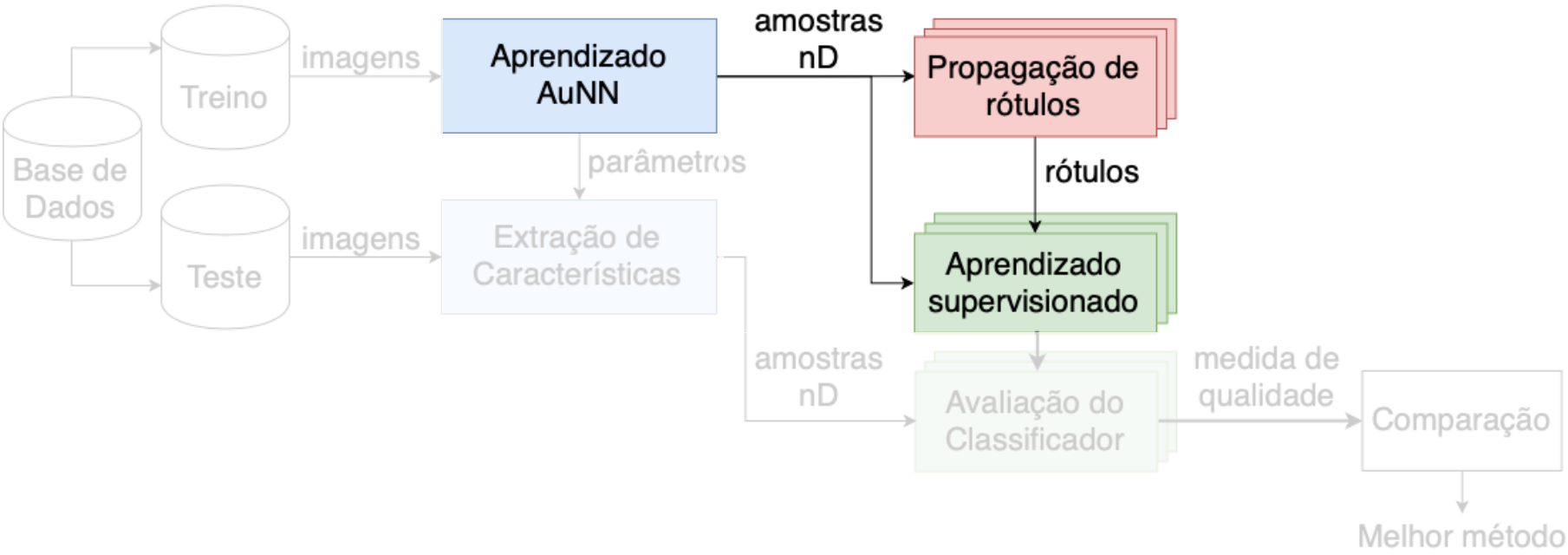


Fig. 6 − Validação do método semi-automático proposto.

Configuração Experimental

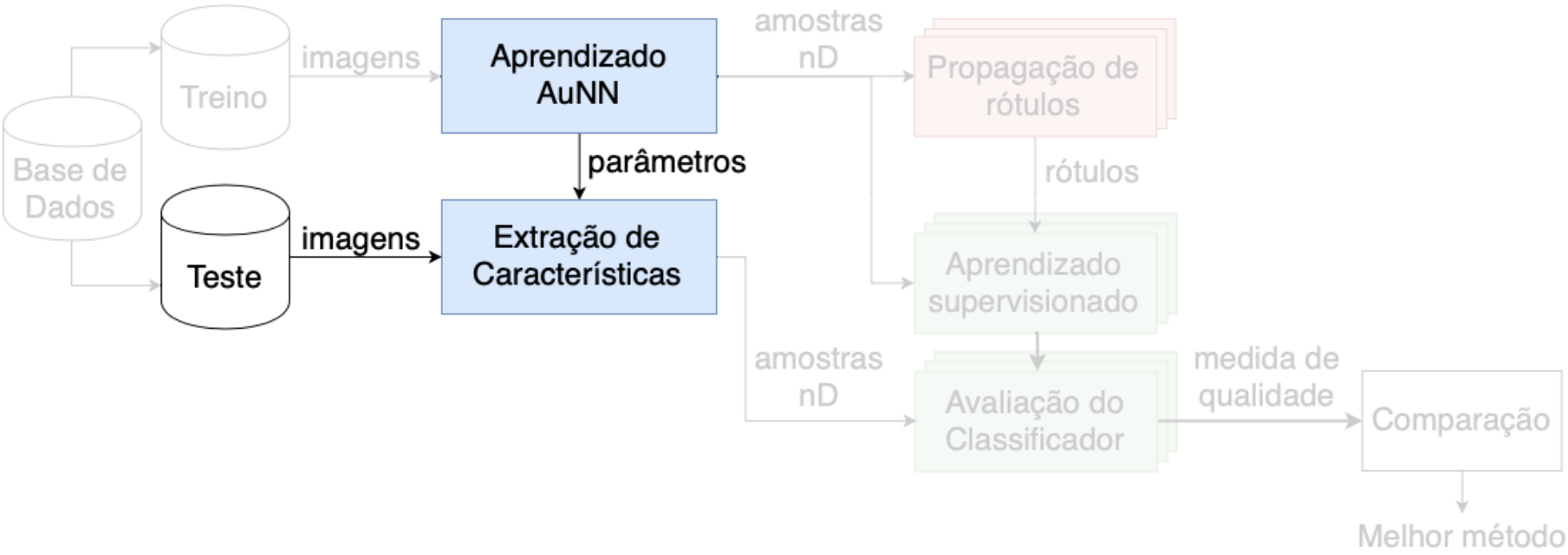


Fig. 6 − Validação do método semi-automático proposto.

Configuração Experimental

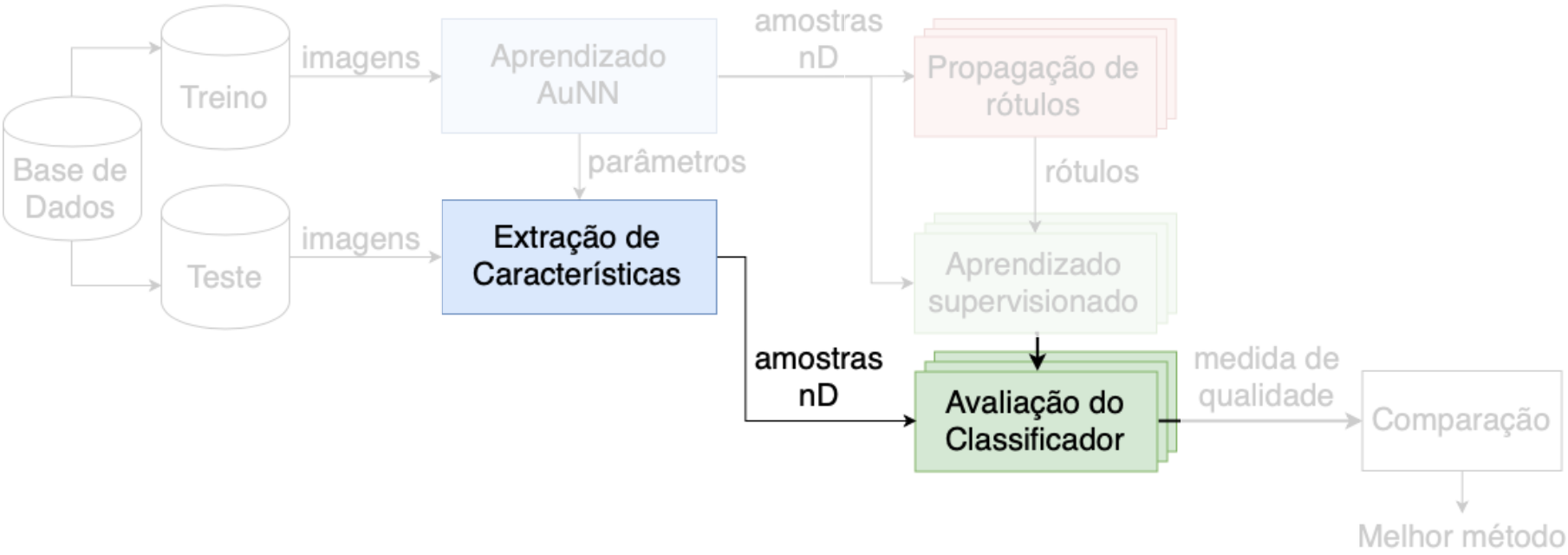


Fig. 6 − Validação do método semi-automático proposto.

Configuração Experimental

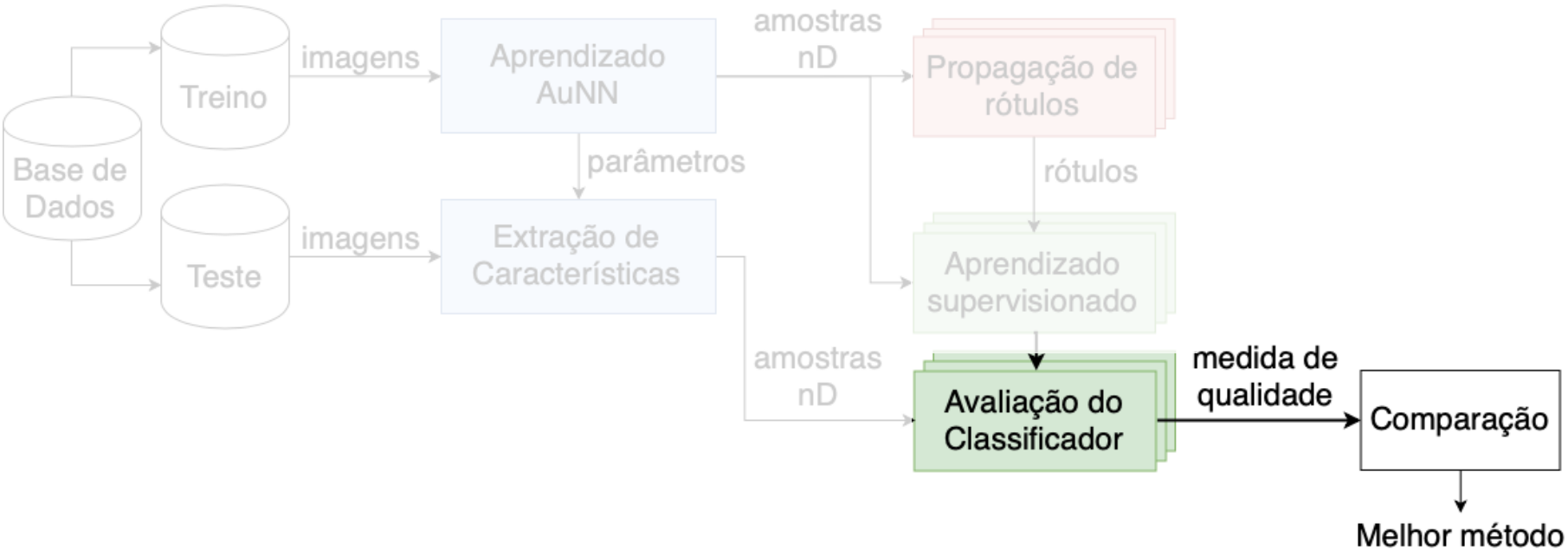


Fig. 6 − Validação do método semi-automático proposto.

Configuração Experimental

3 conjuntos gerados de forma aleatória e divididos em:

- Treinamento: 70%
  - Supervisionado: 3% (S)
  - Não-supervisionado 67% (U) ⇨ Conjunto rotulado
- Teste: 30% (T)

Dimensão do vetor reduzido da AuNN:

- MNIST: 128
- Parasitos: 5.000

Configuração Experimental

Abordagens de propagação de rótulos comparadas:

- Sem propagação
- Automática (ALP): **nD** e **2D**
  - Máquina de Vetores de Suporte Laplaciana (LapSVM)
  - Floresta de Caminhos Ótimos Semi-supervisionada (OPF-Semi)
- Manual: 1 usuário
  - ILP
  - SALP

Classificadores:

- Máquina de Vetores de Suporte (SVM)
- Floresta de Caminhos Ótimos (OPF)

Métrica de qualidade do classificador:

- Cohen Kappa (k)

Experimentos

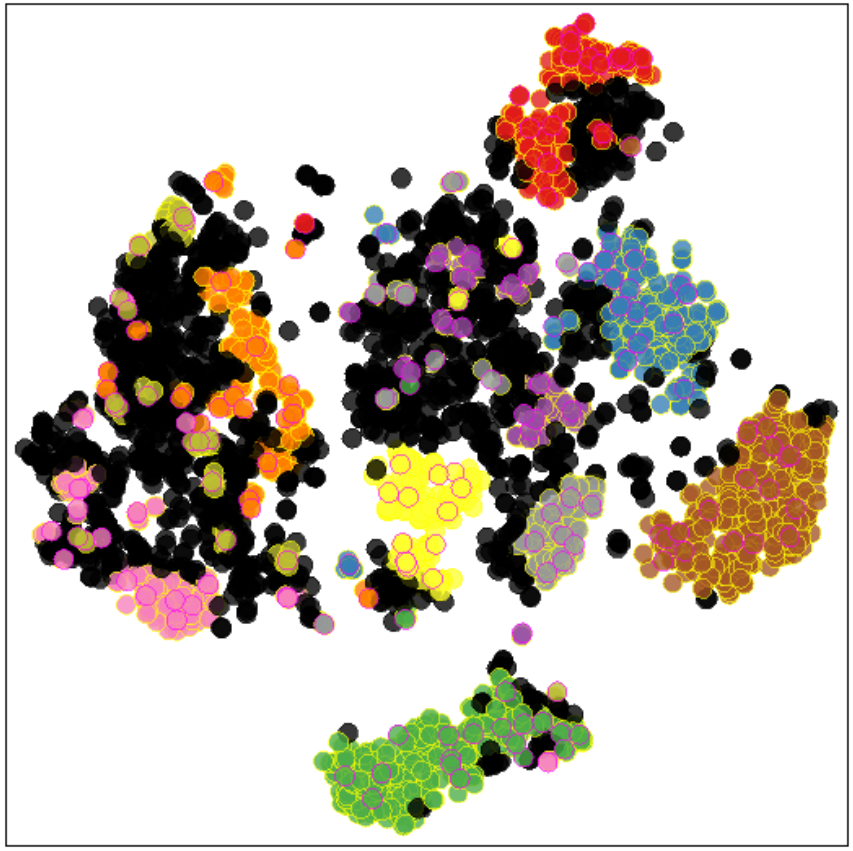
Valores de limiar de confiança foram escolhidos pelo usuário entre 0,5 e 0,6.

- Bases mais fáceis: 0,6
  - Usuário prefere selecionar mais amostras para ILP.

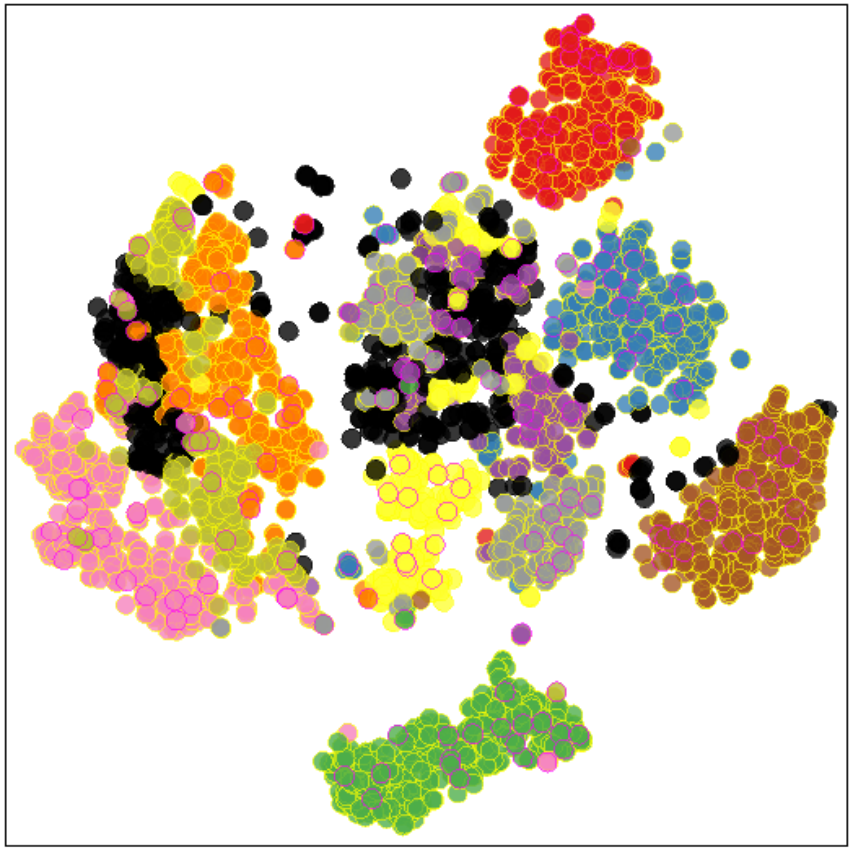
- Bases mais difíceis: 0,5
  - Usuário prefere selecionar mais amostras para ALP.

Experimentos

MNIST (|U|= 3323)



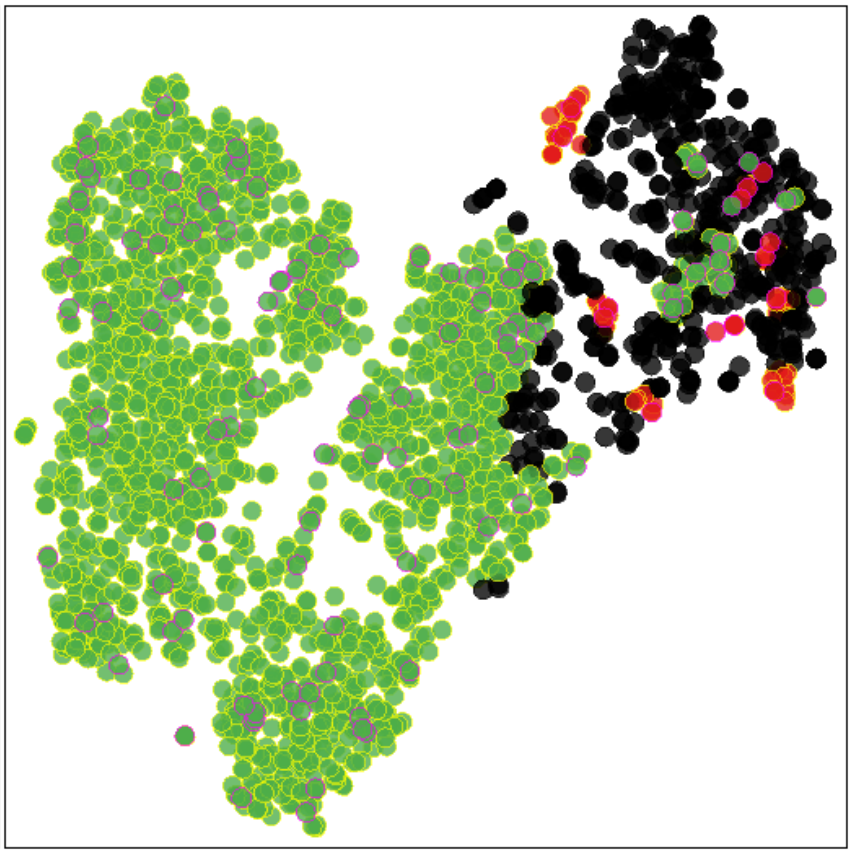
OPF-Semi = 1690



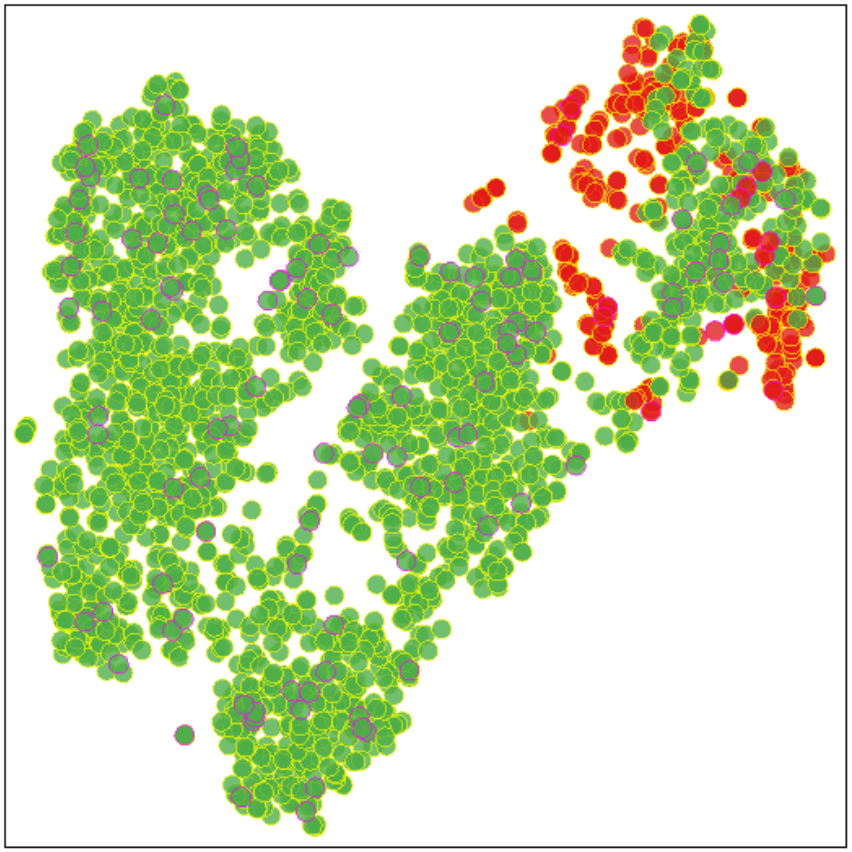
Usuário = 1182

Experimentos

H. Larvae (|U| = 2237)



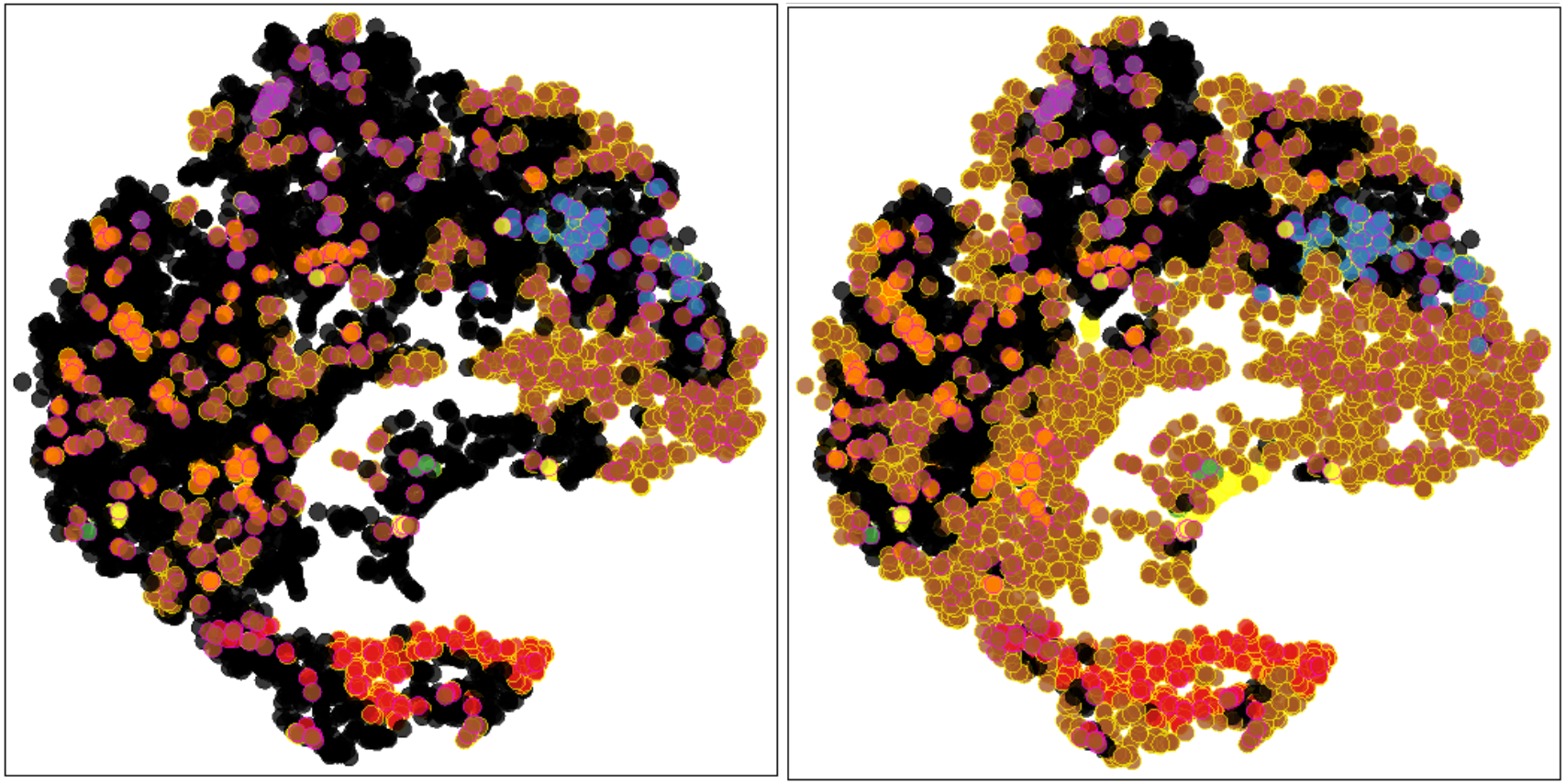
OPF-Semi = 1813



Usuário = 524

Experimentos

Proto. cysts (|U| = 6363)



OPF-Semi = 2217

Usuário = 1733

Experimentos

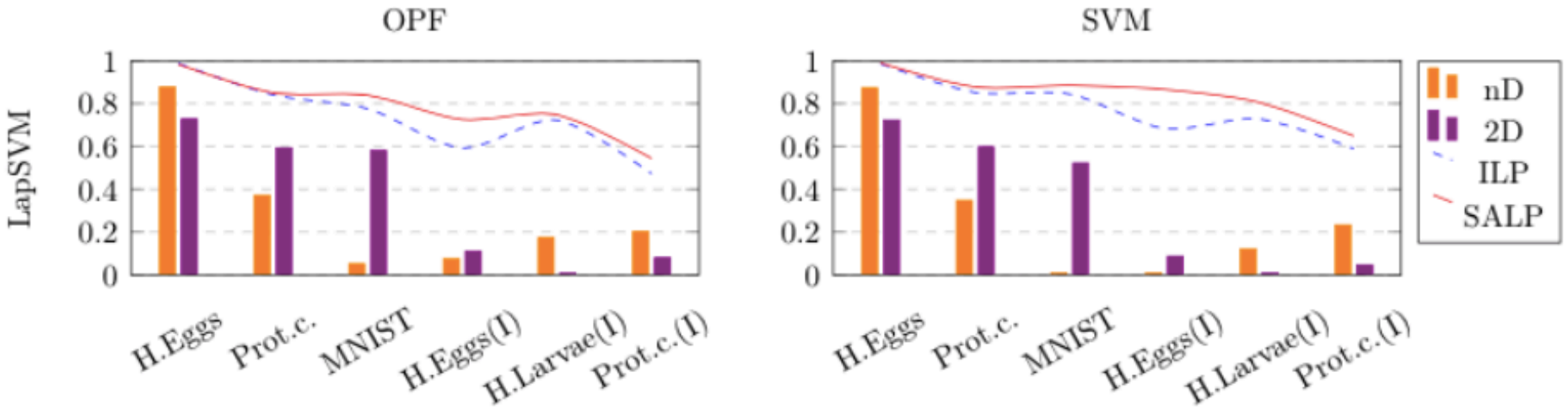


Fig. 7 − Kappa para OPF e SVM para os diferentes métodos automáticos no 2D e ND, além do ILP e SALP.

Experimentos

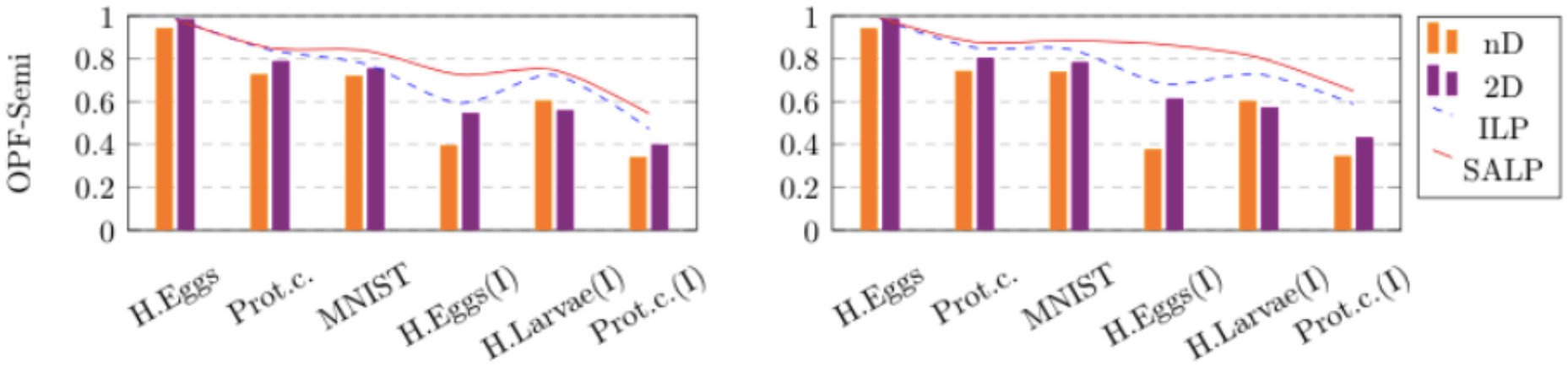


Fig. 8 − Kappa para OPF e SVM para os diferentes métodos automáticos no 2D e ND, além do ILP e SALP.

Experimentos

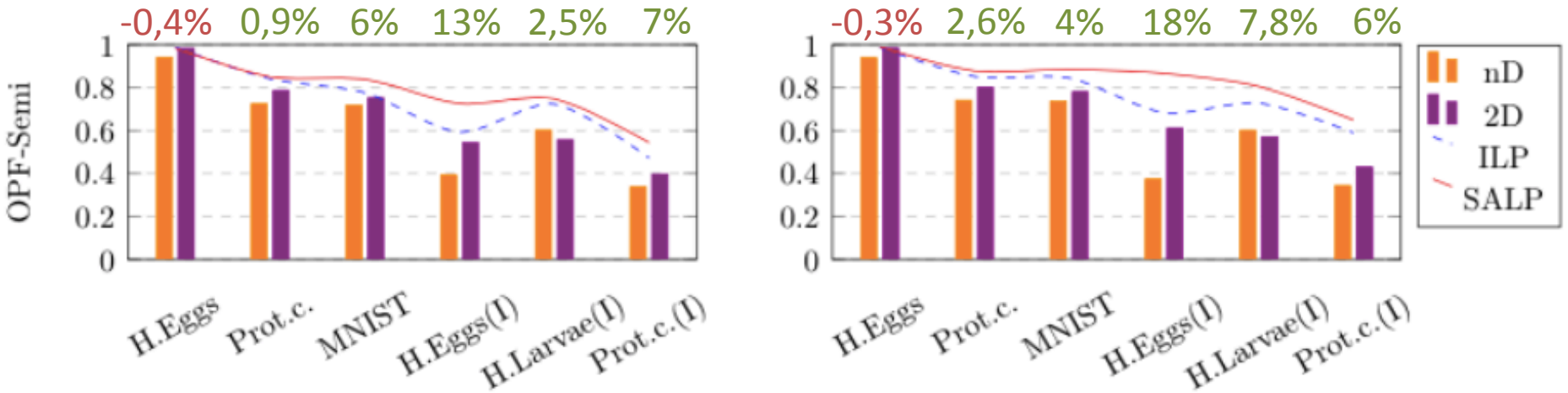


Fig. 8 − Kappa para OPF e SVM para os diferentes métodos automáticos no 2D e ND, além do ILP e SALP.

quantidade de amostras

Redução do esforço do usuário

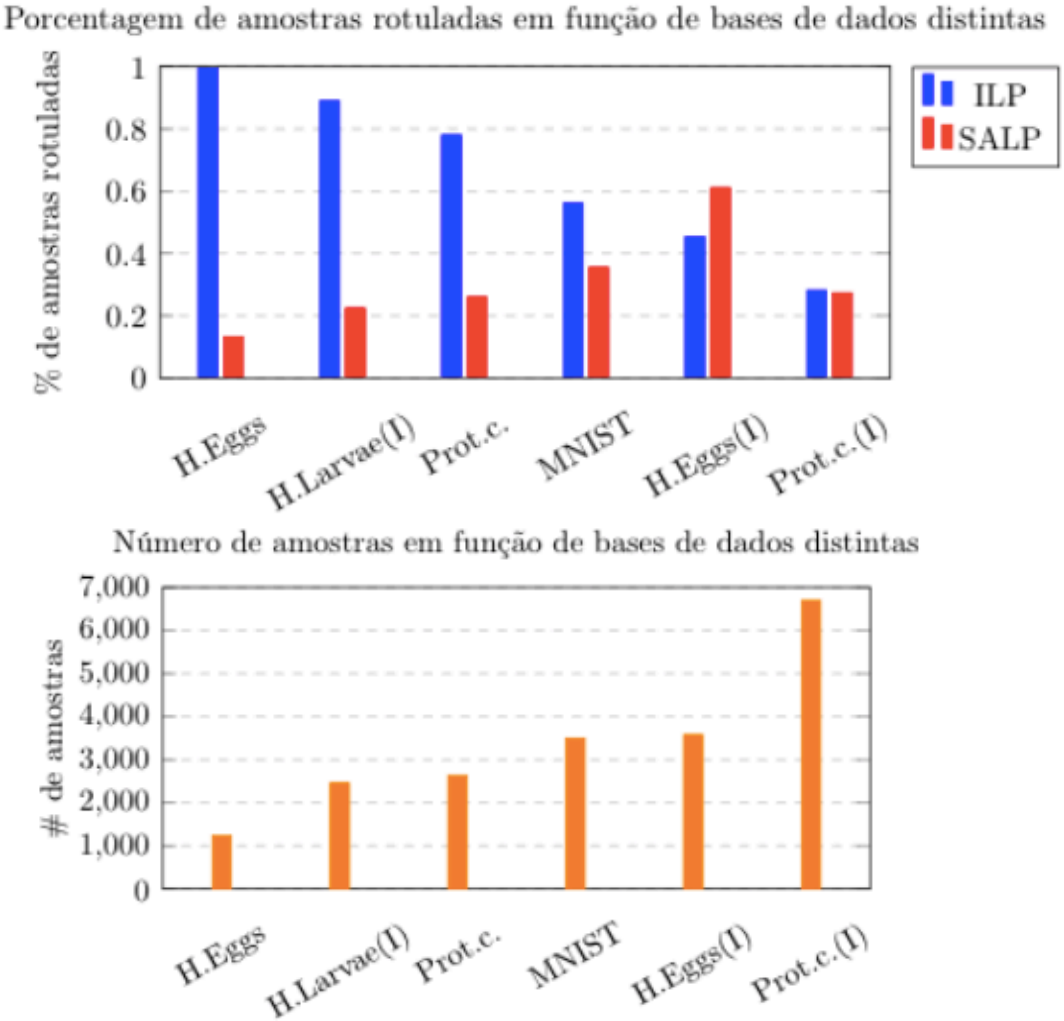


Fig. 9 − Porcentagem de amostras rotuladas e quantidade de amostras nas bases de dados.

Experimentos

Alguns experimentos foram realizados utilizando outra técnica de projeção: UMAP.

- Dificuldade em encontrar parâmetros.
- Resultados piores em termos de acurácia de propagação e classificação (Kappa).
- Padrões que associam os resultados obtidos com o tamanho da base de dados e sua dificuldade.

Discussões

Limitações dos experimentos

- Validação do método:
  - 6 bases de dados, 2 técnicas de classificação e 1 usuário.
- Não foi mensurada de forma quantitativa a qualidade da projeção t-SNE para propagação de rótulos.
  - Métricas
  - Ferramentas visuais

Limitações do método:

- Inerentemente relacionadas às limitações da projeção (cores, muito dado)

Conclusão

Proposta uma abordagem que *combina* o usuário e métodos automáticos para fornecer rótulos para bases de dados insuficientemente anotadas para o treinamento de modelos de classificação.

- AuNN, OPF-Semi, t-SNE e propagação de rótulos interativa.

O espaço projetado no 2D conduz a maiores acurácias de propagação de rótulos automática, do que o espaço latente de alta dimensão extraído pela rede neural.

A abordagem proposta (SALP) alcança melhores resultados de classificação do que a propagação totalmente automática ou totalmente manual, considerando os contextos abordados no presente estudo.

Trabalhos Futuros

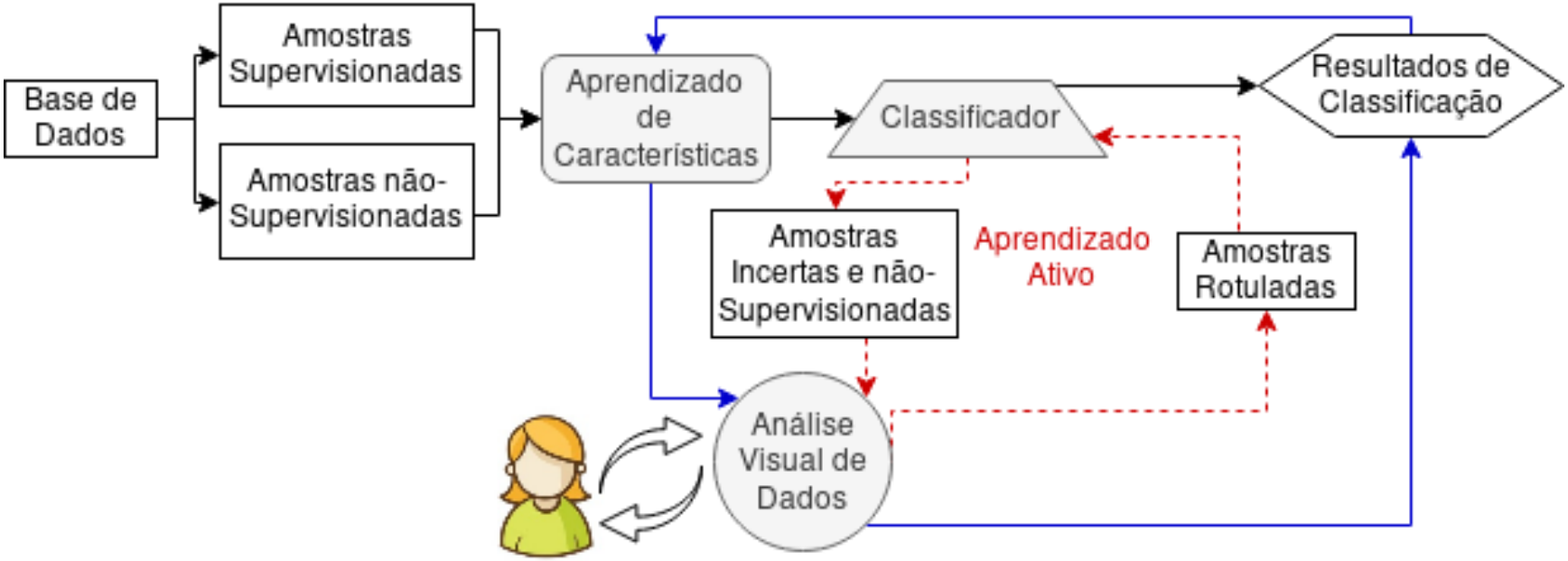


Fig. 10 − Método proposto como trabalho futuro, incluindo o aprendizado ativo.