

TabEBM: A Tabular Data Augmentation Method with Distinct Class-Specific Energy-Based Models

Andrei Margeloiu^{1*}, Xiangjian Jiang^{1*}, Nikola Simidjievski^{2,1}, Mateja Jamnik¹

¹Department of Computer Science and Technology, University of Cambridge, UK

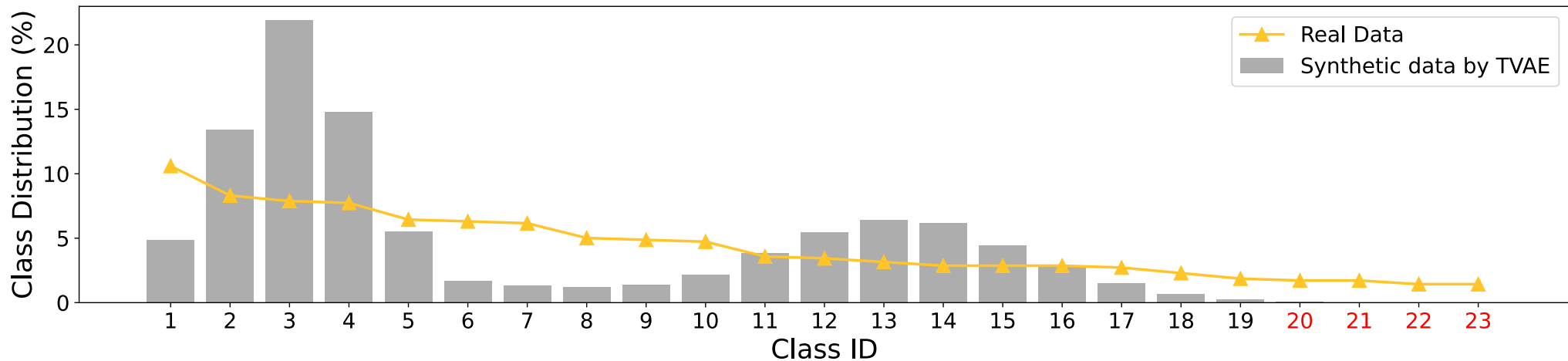
²PBCI, Department of Oncology, University of Cambridge, UK

{am2770, xj265, ns779, mj201}@cam.ac.uk

December, 2024

Tabular data augmentation is challenging

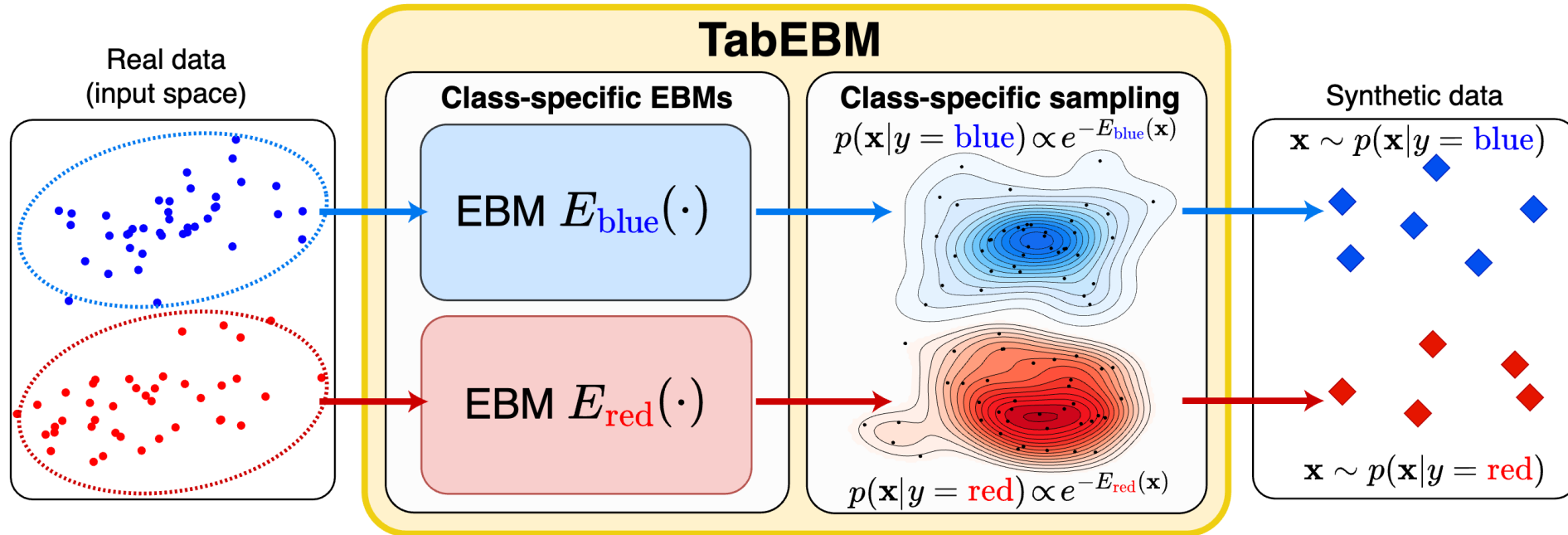
- Challenges in tabular data
 - intricate acquisition
 - lack of symmetries
- Observations: Generative tabular augmentation methods tend to have
 - poor utility
 - mode collapse
 - failure to generate for specific classes



Ideal augmentation should be class-specific

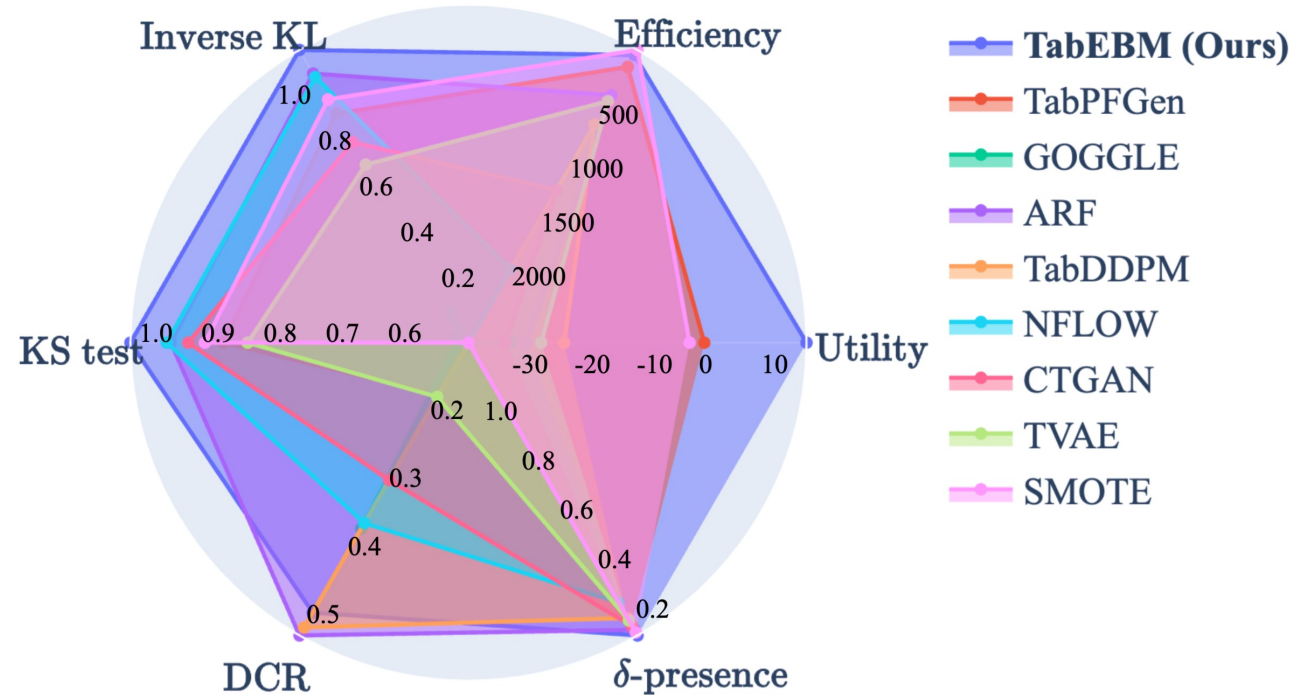
- **Reduced negative impacts across classes**
 - generate samples free from noise and data imbalance induced by other classes
- **Flexible generation for any class**
 - maintain the original stratification in real data
- **High practicability (unlimited class numbers)**
 - handle any datasets without limitations from number of classes

TabEBM performs class-specific augmentation



- **Step 1:** For each class c , fit a pretrained classifier f_{θ}^c on a **surrogate binary classification task**, which is **training-free**
- **Step 2:** Reinterpret its logits into **class-specific energy**
- **Step 3:** Augment class c with data generated via SGLD **sampling on the class-specific energy surface**

TabEBM generates high-quality synthetic data



TabEBM is a performant **training-free** tabular data augmentation method, ready to use with **only a few lines of code**.

```

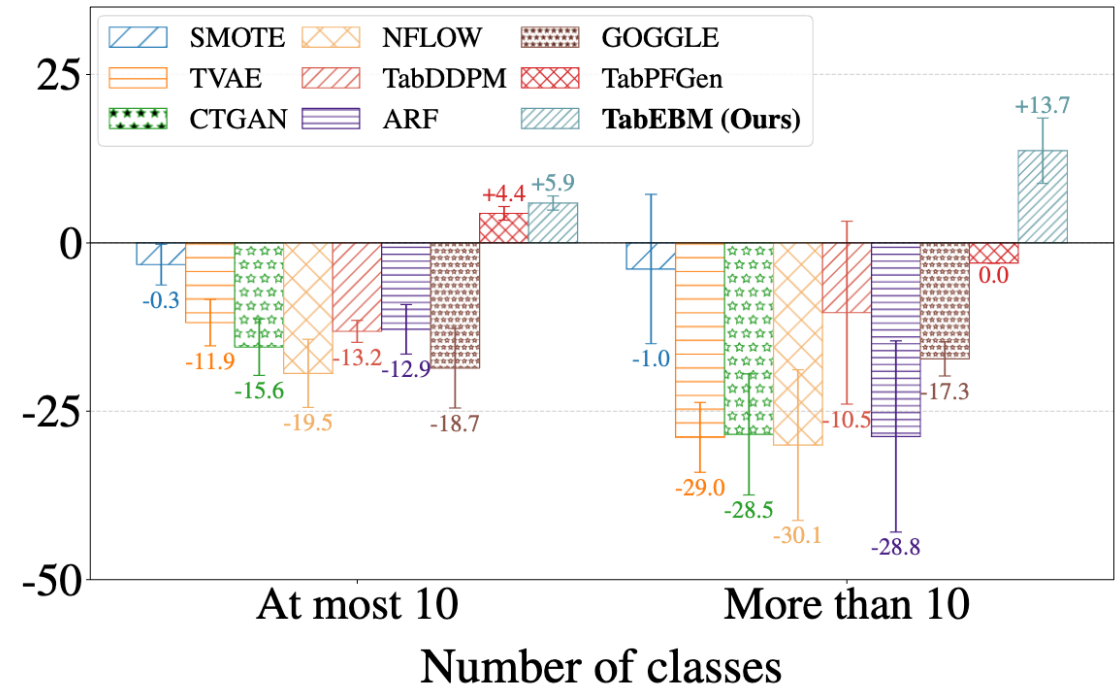
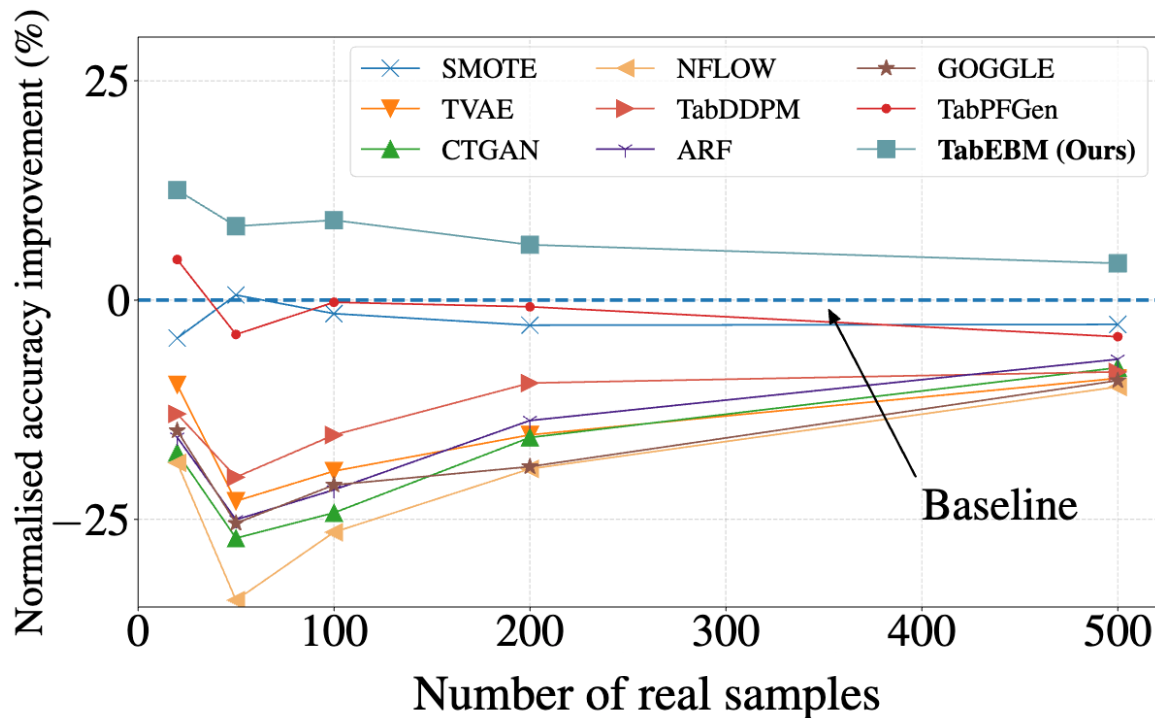
1 from tabebm.TabEBM import TabEBM
2
3 tabebm = TabEBM()
4 augmented_data = tabebm.generate(X_train, y_train, num_samples=100)

```



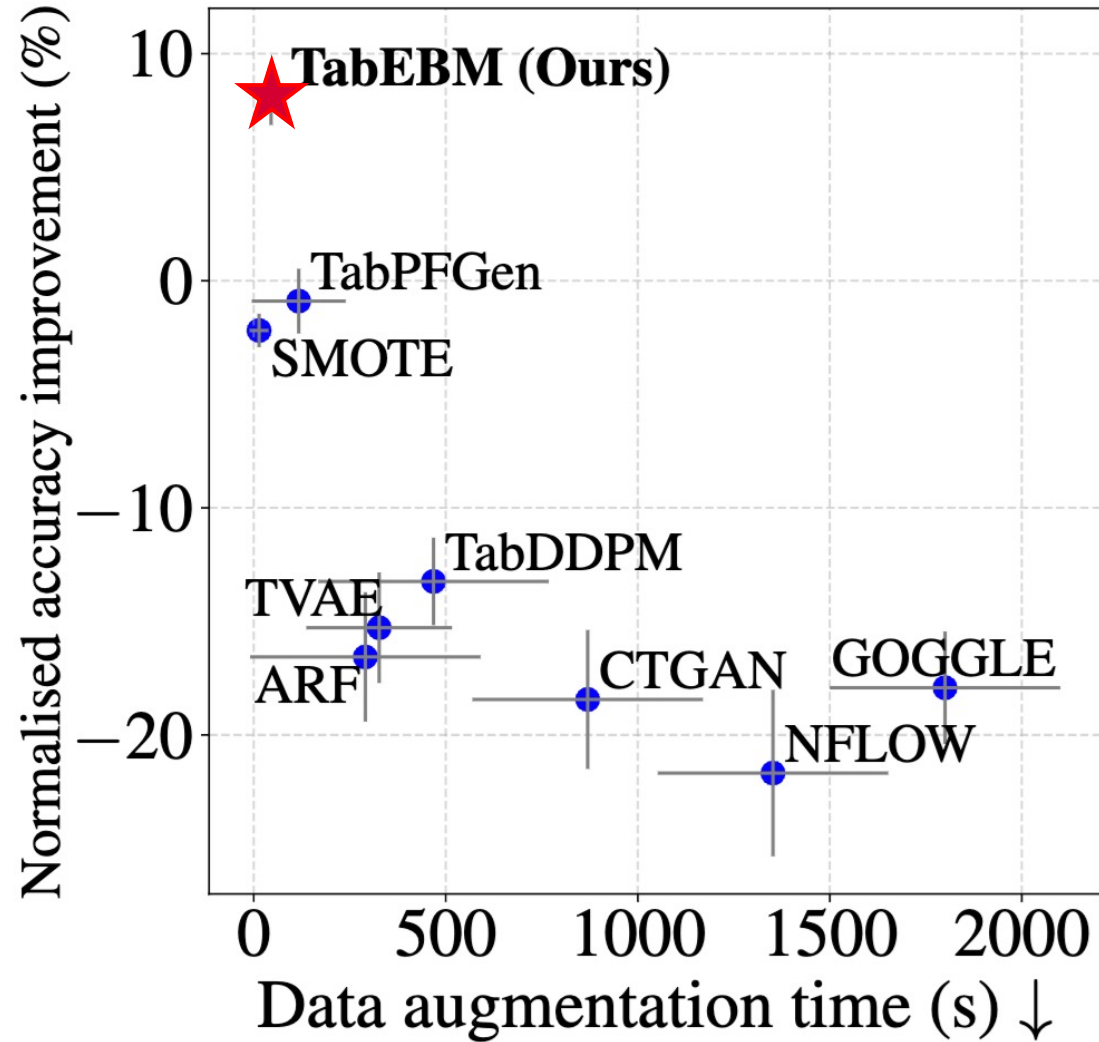
Code

TabEBM improves downstream accuracy



Mean normalised balanced accuracy improvement (%) across 33 test cases of different sample sizes (Left) and varying numbers of classes (Right).

TabEBM is computationally efficient




Median data augmentation time vs. mean normalised balanced accuracy.

Thanks

For more details, please refer to
our paper and code!

Reach out via {am2770, xj265}@cam.ac.uk 🥳



 Paper
(ID: 12221)



 Code