AFRO-MNIST: SYNTHETIC GENERATION OF MNIST-STYLE DATASETS FOR LOW-RESOURCE LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

We present Afro-MNIST, a set of synthetic MNIST-style datasets for four orthographies used in Afro-Asiatic and Niger-Congo languages: Ge‘ez (Ethiopic), Vai, Osmanya, and N'Ko. These datasets serve as "drop-in" replacements for MNIST. We also describe and open-source a method for synthetic MNIST-style dataset generation from single examples of each digit. These datasets can be found at `https://github.com/ANONYMOUS_AUTHOR/AfroMNIST`. We hope that MNIST-style datasets will be developed for other numeral systems, and that these datasets vitalize machine learning education in underrepresented nations in the research community.

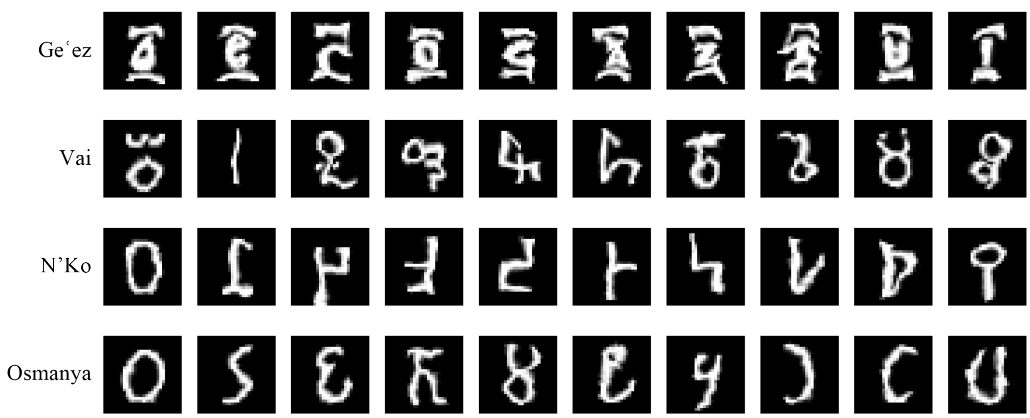

Figure 1: Afro-MNIST: Four synthetic MNIST-style datasets.

1 MOTIVATION

Classifying MNIST Hindu-Arabic numerals (LeCun et al. (1998)) has become the "Hello World!" challenge of the machine learning community. This task has excited both a large number of prospective machine learning scientists, and led to practical advancements in optical character recognition.

The Hindu-Arabic numeral system is the predominant numeral system used in the world today. However, there are a sizeable number of languages where the numeric glyphs are not inherited from the Hindu-Arabic numeral system. As it relates to human language, work in machine learning and artificial intelligence focuses almost exclusively on high-resource languages such as English and Mandarin Chinese.

Unfortunately, "mainstream" languages such as English and Chinese comprise only a tiny fraction of the world's languages — of over 7,000 languages in the world, the vast majority are not represented in the machine learning research community. In particular, we note that there is a vast collection of alternative numeral systems, for which an MNIST-style dataset is not available.

In an effort to make machine learning education more accessible to diverse groups of people, it is imperative that we develop datasets which represent the heterogeneity of existing numeral systems.

Furthermore, as studied in Mgqwashu (2011), familiarity of the glyph shapes from the students' mother tongue facilitates learning and enhanced epistemological access. With the recent drive towards spreading AI literacy in Africa, we argue that it would be rather useful to use local numeral glyphs in the initial *Machine Learning 101* courses to spark both enthusiasm and establish familiarity.

Additionally, we were inspired by the dire warning that appears in the work by Mgqwashu (2011) that *the numeral system is the most endangered aspect of any language* and that the excuse of not having these datasets and downstream OCR applications will further accelerate the death of several native numeral systems dealt with in this work.

In similar work, Prabhu (2019) collected and open-sourced an MNIST-style dataset for the Kannada language; however, this effort took a significant amount of manpower from 65 volunteers. While this methodology is effective for high-resource languages, we expect that it is not practical to apply it for the large number of low-resource languages.

Much of the world's linguistic diversity comes from under-resourced languages spoken in developing nations. In particular, there is a wealth of linguistic diversity in the languages of Africa, many of which have dedicated orthographies and numeral systems. In this work, we focus on the Ge'ez (Ethiopic), Vai, Osmanya, and N'Ko scripts[1], each of which have dedicated digits. In particular, the Ge'ez script is used to transcribe languages such as Amharic and Tigrinya, which are spoken by some 30 million people (Eberhard et al. (2019)).

Because large amounts of training data for African languages such as Amharic and Somali are not readily available, we experiment with creating synthetic numerals that mimic the likeness of handwritten numerals in their writing systems. Previous work in few-shot learning and representation learning has shown that effective neural networks can be trained on highly perturbed versions of even just a single image of each class (Dosovitskiy et al. (2015)). Thus, we propose the synthetic generation of MNIST-style datasets, from Unicode exemplars of each numeral.

Our Contributions:

1. We release synthetic MNIST-style datasets for four Afro-Asiatic or Niger-Congo languages: Ge'ez, Vai, Osmanya, N'Ko.
2. We describe a general framework for manpower-light synthesis of MNIST-style datasets.
3. We explore the impact of orthographies on the effectiveness of machine learning models.

2 METHODOLOGY

Inspired by the work in Prabhu et al. (2019), we first generate an exemplar *seed* dataset for each numeral system (Figure 2) from the corresponding Unicode characters. From a category theoretic perspective, these would constitute the classwise prototypes as they (following Qi et al. (2006))

1. Reflect the central tendency of the instances' properties or patterns;
2. Are more similar to some category members than others;
3. Are themself self-realizable but are not necessarily an instance.

In order to generate synthetic examples from these prototypical glyphs, we apply elastic deformations and corruptions (similar to Mu & Gilmer (2019) and as mentioned in Simard et al. (2003)) to the exemplars. We empirically chose elastic deformation parameters $\alpha = 8, \gamma \in [2, 2.5]$ in order to maximize variance while still retaining visual distinguishability. The impact of these parameters on the synthetic image is explored in the Appendix.

We compared our result to a small dataset of written Ge'ez digits (Molla (2019)); the differences in exemplar versus handwritten generated data are shown in Figure 3. We note that in cases where a limited amount of handwritten digit data is available, deformations and corruptions can be applied to those examples instead of Unicode exemplars.

[1]We note that the Vai, Osmanya, and N'Ko scripts are not in wide use, but nonetheless they can be synthesized using the methods we present.

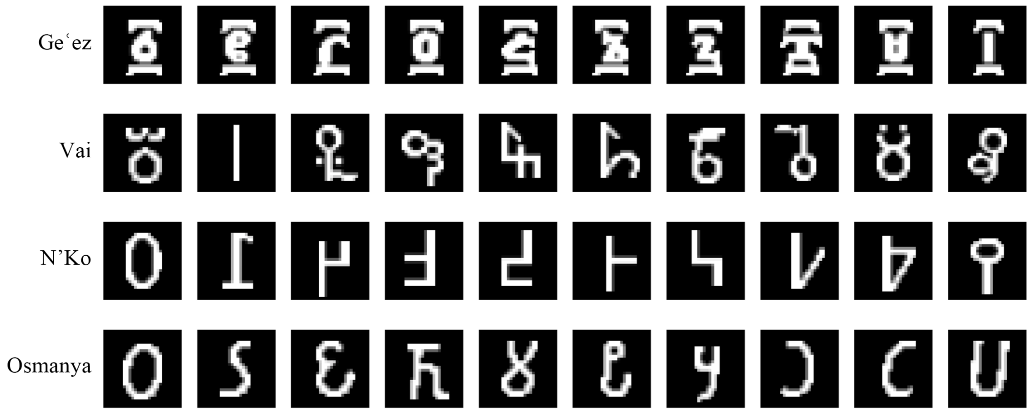

Figure 2: Exemplar images generated from Unicode characters

Figure 3: Comparison of Ge‘ez-MNIST generated from exemplars and handwritten samples.

3 Dataset Details

To produce “drop-in” replacements for MNIST, we closely emulate the format of the latter. Each of our datasets contains with 60000 training images and 10000 testing images. Each image is greyscale, and 28×28 pixels in size. Each dataset contains an equal number of images of the numerals zero to nine[2]. To assess the robustness of our synthetic generation pipeline, we plot the global average of each character (Figure 4) and note that it closely aligns with our exemplars.

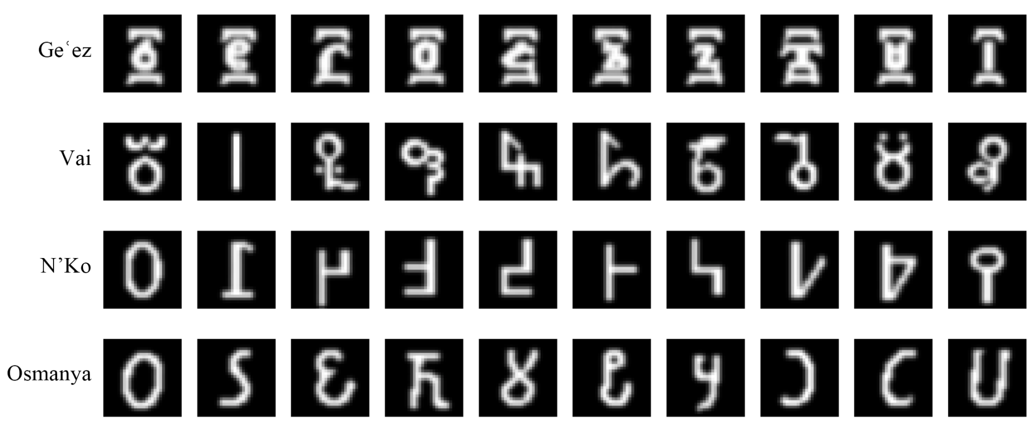

Figure 4: Global average of each glyph in AfroMNIST datasets.

We are also interested in the morphological differences between our generated datasets and the original (Hindu-Arabic) MNIST. We visualize the morphological characteristics of our datasets ac-

[2]The Ge‘ez script lacks the digit 0, so our classes represent the numerals 1-10 for that script.

cording to the methodology of Castro et al. (2019), and also plot the UMAP embeddings of the data. These analyses on Ge‘ez-MNIST are shown in Figures 5 and 6. Analogous analyses for the other datasets are given in the Appendix.

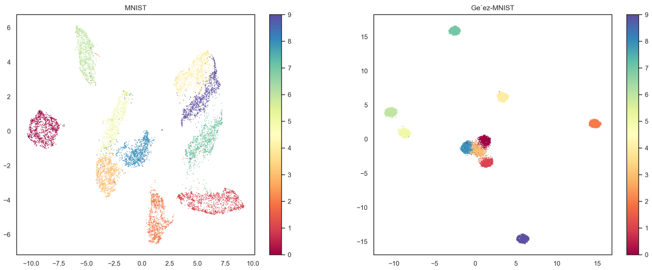

Figure 5: Comparison of UMAPs of MNIST and Ge‘ez-MNIST.

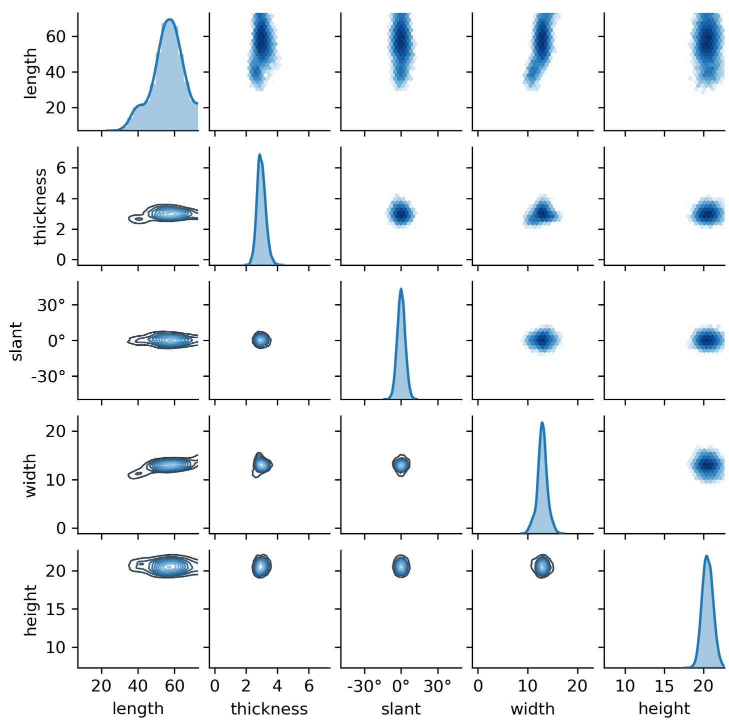

Figure 6: Morphological comparison of MNIST versus Ge‘ez-MNIST

4 EXPERIMENTS

In order to provide a baseline for machine learning methods on these datasets, we train LeNet-5 (LeCun et al. (1998)), the network first used on the original MNIST dataset, on our datasets. We train with the Adam optimizer, with an initial learning rate of 0.001, on the categorical crossentropy loss. Models were trained until convergence. The model architecture is described in the Appendix. The results are shown in Table 1.

We note that certain scripts have a comparatively high variability when written as opposed to our exemplars, and this leads to poor generalization (Figure 3, 6). Furthermore, it's clear from examination of UMAP embeddings that our dataset is not as heterogenous as in-the-wild handwriting. We tested a LeNet-5 trained on Ge‘ez-MNIST on the aforementioned dataset of handwritten Ge‘ez

Dataset	Accuracy (%)
MNIST	99.65
Ge‘ez-MNIST	99.92
Vai-MNIST	100
Osmanya-MNIST	99.99
N’Ko-MNIST	100

Table 1: Accuracy of LeNet trained on our datasets

numerals, and found the model achieved only an accuracy of 30.30%. Certain numerals were easier to distinguish than others (Figure 7). We expect this benchmark to be a fertile starting-point for exploring augmentation and transfer learning strategies for low-resource languages.

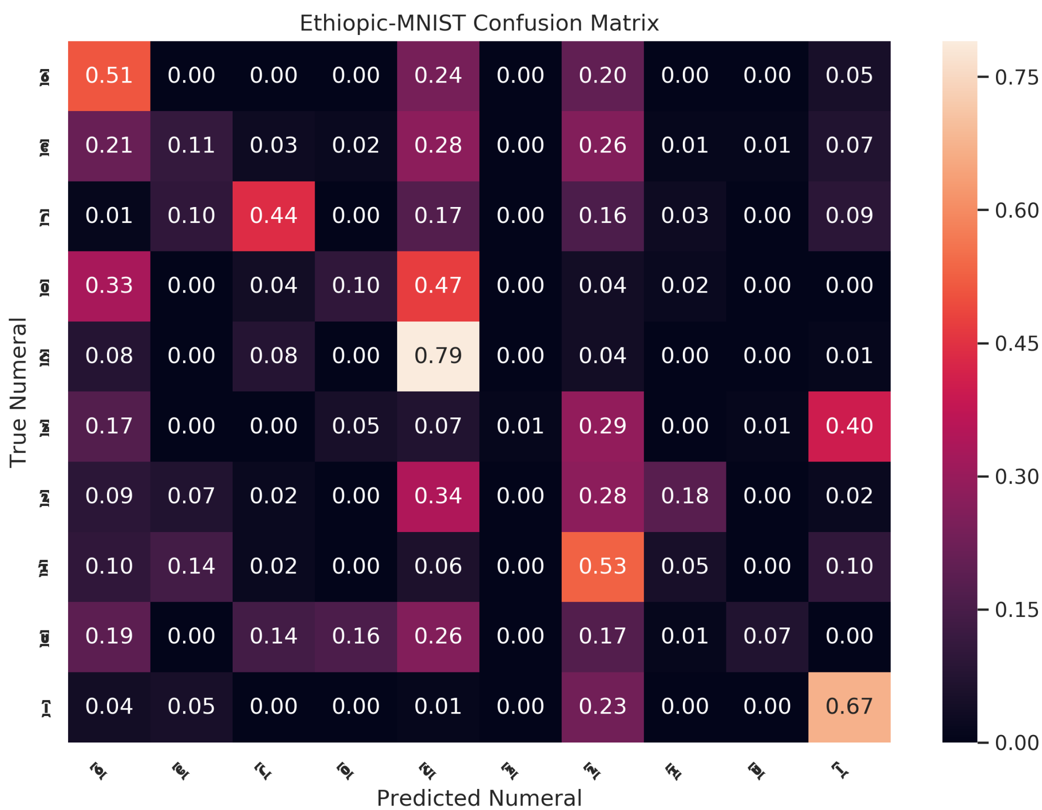

Figure 7: Confusion Matrix of LeNet-5 trained on the synthetic Ge‘ez-MNIST, tested on the hand-written dataset.

5 Conclusion

We present AfroMNIST — a set of MNIST-style datasets for numerals in four African orthographies. It is our hope that the availability of these datasets enables the next generation of diverse research scientists to have their own ”Hello World” moment. We also open-source a simple pipeline to generate these datasets without any manual data collection, and we look forward to seeing the release of a wide range of new MNIST-style challenges from the machine learning community for other numeral systems.

REFERENCES

Daniel C. Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-MNIST: Quantitative assessment and diagnostics for representation learning. *Journal of Machine Learning Research*, 20, 2019.

Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015.

David M. Eberhard, Gary F Simons, and Charles D Fennig. Languages of the world, 2019. URL `http://www.ethnologue.com/`.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Emmanuel Mfanafuthi Mgqwashu. Academic literacy in the mother tongue: A pre-requisite for epistemological access. *Diversity, Transformation and Student Experience in Higher Education Teaching and Learning*, pp. 159, 2011.

Tesfamichael Molla. Ethiopian-mnist. `https://github.com/Tesfamichael1074/Ethiopian-MNIST`, 2019.

Norman Mu and Justin Gilmer. Mnist-c: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337*, 2019.

Vinay Uday Prabhu. Kannada-mnist: A new handwritten digits dataset for the kannada language. *arXiv preprint arXiv:1908.01242*, 2019.

Vinay Uday Prabhu, Sanghyun Han, Dian Ang Yap, Mihail Douhaniaris, Preethi Seshadri, and John Whaley. Fonts-2-handwriting: A seed-augment-train framework for universal digit classification. *arXiv preprint arXiv:1905.08633*, 2019.

Feng Qi, A-Xing Zhu, Mark Harrower, and James E Burt. Fuzzy soil mapping based on prototype category theory. *Geoderma*, 136(3-4):774–787, 2006.

Patrice Y Simard, David Steinkraus, John C Platt, et al. Best practices for convolutional neural networks applied to visual document analysis. In *Icdar*, volume 3, 2003.

A APPENDIX

Layer	Filter Size	Units	Parameters
Conv2D	5x5	6	156
Max Pool	2x2	6	0
Tanh			0
Conv2D	5x5	16	2416
Max Pool	2x2	6	0
Tanh		0	
Flatten			
Dense		120	30804
Tanh			0
Dense		84	10164
Tanh			0
Dense		10	850
Softmax			

Table 2: The architecture of LeNet from LeCun et al. (1998)

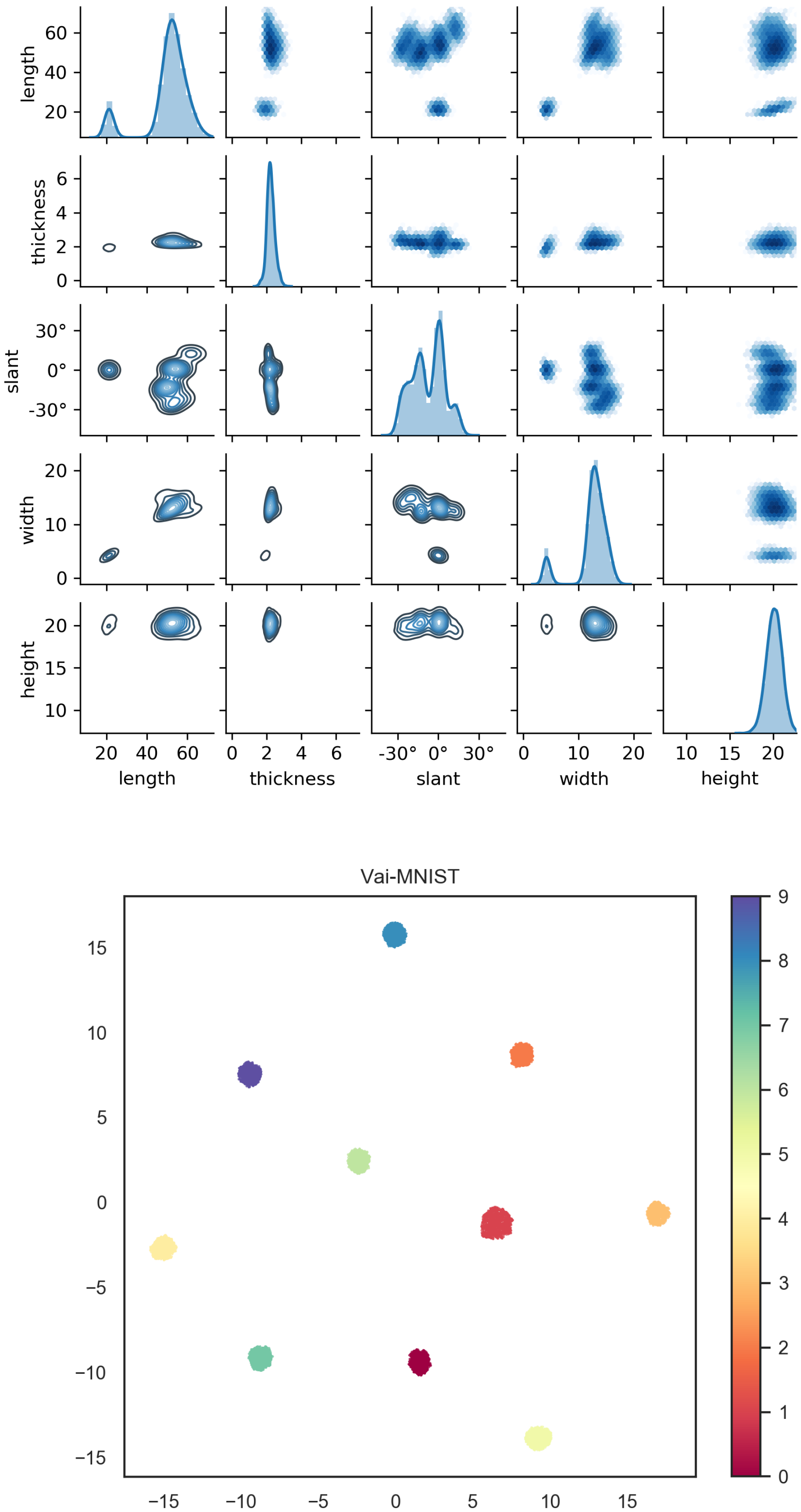

Figure 8: Vai-MNIST morphology and UMAP embedding.

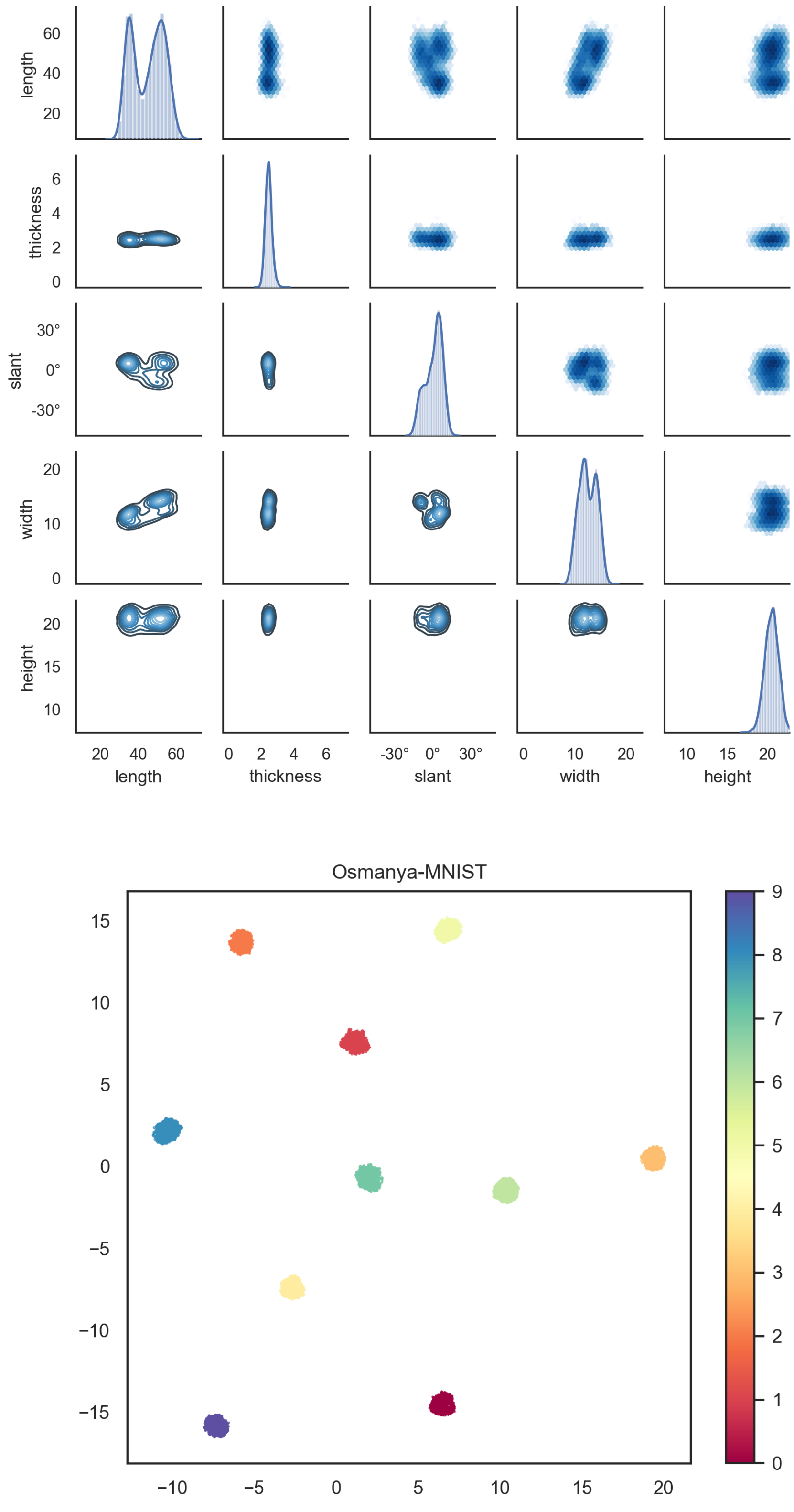

Figure 9: Osmanya-MNIST morphology and UMAP embedding.

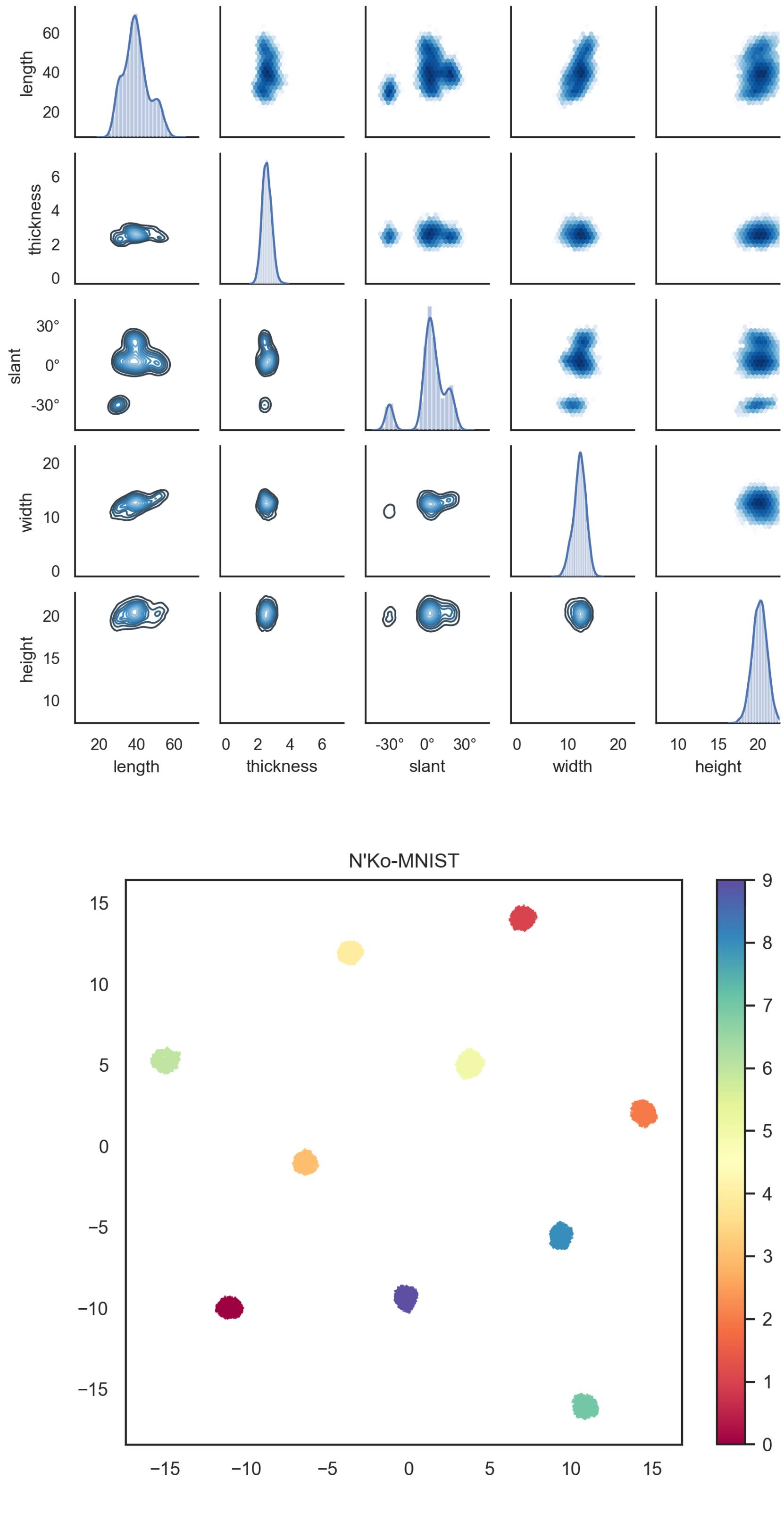

Figure 10: N’Ko-MNIST morphology and UMAP embedding.

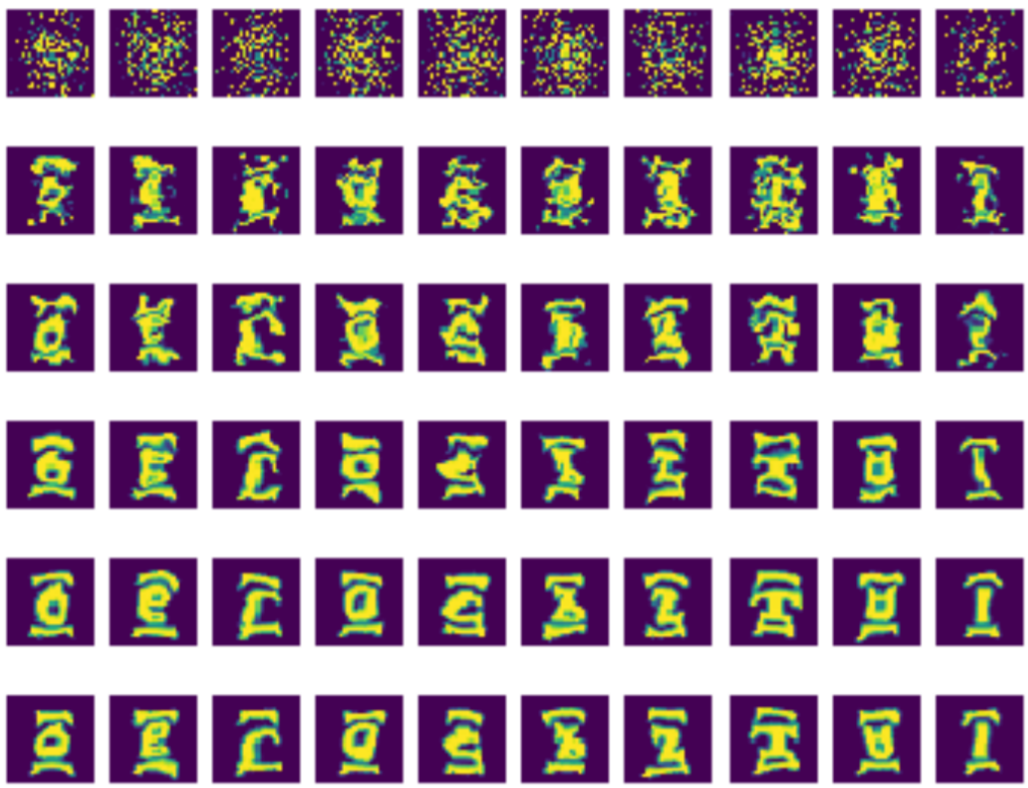

Figure 11: Impact of γ on generated Ge'ez numerals. From top to bottom, $\gamma = 0.01, 1, 1.5, 2, 2.5, 3$