Набор данных рукописных символов русского языка.

Д. В.Яценко

ФВТ, каф. системного анализа, Южный Федеральный Университет,
Мильчакова 10, Ростов-на-Дону 344090, Россия
d.yacenko@gmail.com

К. В.Смирнов

Университет Иннополис,
ул. Университетская, д. 1, Верхнеуслонский район 420500, Татарстан,Россия
k.smirnov@innopolis.university

1 мая 2024 г.

УДК *004.08*

Аннотация

Наличие наборов данных - важный фактор для развития машинного обучения. Сегодня опубликовано не так много наборов данных, содержащих национальные рукописные символы. В статье представлен набор данных, описывающий рукописные символы русского языка. Датасет содержит образцы написания символов 12 писателей и содержит прописные и строчные символы, а также цифры - итого 76 классов символов. Дополнительно набор данных включает некоторые образцы написания целых слов. Данные о символах представлены в виде координат траекторий движения пера писателя в процессе начертания символа. В работе описывается методология сбора и обработки данных. Кроме того, сравнивается точность распознавания базовой моделью LeNet-5 представленного набора данных, с близким по количеству классов набором данных EMNIST. Для представленного датасета демонстрируется качество немного хуже, чем для EMNIST. Однако датасет, при этом обладает важными свойствами - поскольку данные заданы не только матричным видом , но и траекторией движения - это позволяет учитывать особенности движения пера, что может быть использовано не только в целях классификации символов, но и с целью распознавания персональных особенностей почерка, переноса стиля письма, графологического анализа.

Ключевые слова

набор данных, нейронная сеть, рукописные символы, траектория написания.

1 Введение

Развитие области машинного обучения в немалой степени обеспечивается различными наборами данных. Публикация таких наборов данных позволяет получить независимый способ анализа и сравнительной оценки различных подходов и алгоритмов обучения. Это позволяет исследователям быстро получить представление о производительности и особенностях методов и алгоритмов.

Однако каждый конкретный набор данных обеспечивает решение узкого класса задач, поэтому появление широкого спектра датасетов и эталонных задач важно для обеспечения более целостного подхода к оценке и характеристике производительности алгоритмов и моделей машинного обучения. Сегодня опубликовано несколько стандартизированных наборов данных, которые широко используются и стали весьма конкурентоспособными (таб 1).

Таблица 1: Опубликованные наборы данных символов различных языков.

Название датасета	**Краткое описание**	**Препроцессинг**	**Размер**	**Формат**
Artificial Characters Dataset	10 заглавных символов англ языка [1].	Координаты линий. Другие функции.	6000	Текст
Letter Dataset	заглавные печатные английские символы [2, 3].	17 признаков, выделенных из всех символов.	20000	Текст
CASIA-HWDB	рукописные китайские символы, 3755 классов [4].	черно-белые символы	1172907	Изображения, текст
CASIA-OLHWDB	рукописные китайские символы, 3755 классов [5, 5].	последовательность координат точек	1174364	Изображения, текст
Character Trajectories Dataset	простые символы [6, 7].	трехмерная матрица траектории скорости кончика пера	2858	Текст
Chars74K Dataset	Естественные изображения символов, исп. в Англии и в Канаде [8].		74107	
EMNIST dataset	рукописные символы от 3600 участников [9].	Создано в продолжение NIST. Изображения 28x28, соответствующие набору MNIST.	800000	Изображения
UJI Pen Characters Dataset	Изолированные рукописные символы [10, 11].	координаты положения пера.	11640	Текст
Gisette Dataset	Образцы почерка из часто путаемых 4 и 9 символов [12].	Данные, извлеченные из изображений рукописного ввода.	13500	Изображения, текст
Omniglot dataset	1623 различные рукописные символы из 50 различных алфавитов [13].	ручная маркировка.	38300	Изображения, текст
MNIST database	База рукописных символов [14, 15].	ручная маркировка .	60000	Изображения, текст
Optical Recognition of Handwritten Digits Dataset	Нормализованные растровые изображения рукописных данных [16].	Нормализованные размеры	5620	Изображения, текст
Pen-Based Recognition of Handwritten Digits Dataset	Рукописные цифры на электронном планшете [17, 18].		10992	Изображения, текст
Semeion Handwritten Digit Dataset	Рукописные цифры от 80 человек [19].	Все рукописные цифры были нормализованы по размеру.	1593	Изображения, текст
HASYv2	Рукописные математические символы [20].	Символы отцентрированы и имеют размер 32x32.	168233	Изображения, текст
Noisy Handwritten Bangla Dataset	Рукописные числа и символы, также имеется три типа шума [21, 22].	Символы отцентрированы и имеют размер 32x32..	99330	Изображения, текст

Перечисленные в таблице (таб 1) датасеты не покрывают полностью области исследований распознавания символов. В частности:

- существенная часть символов – печатные или рукописные, но в ”печатном”стиле,
- большинство наборов данных предоставляют изображения цифр и/или английских букв,
- наборов данных символов национальных языков очень мало,
- подавляющее большинство предоставляют в качестве данных растровое изображение символов, без данных о траектории написания символов.

Однако в некоторых задачах обработки рукописных символов траектория начертания символа является важным информативным признаком. Это, например, следующие классы задач:

- распознавание символов в процессе написания, например, на интерактивных планшетах,
- распознавание символов плохого почерка, например, людей с ограниченными возможностями или травмами конечностей,
- распознавание авторских особенностей почерка и распознавание почерка,
- распознавание подписей, вензелей, као,
- перенос стиля почерка,
- графологический анализ текстов.

Анализ опубликованных наборов данных выявил отсутствие русскоязычного полносимвольного набора данных с информацией о траектории начертания символов.

В этой статье представлен такой набор наборов данных, опубликованный под названием Russian Handwritings Tracked [23]. Далее в статье описываются характеристики датасета, документирован процесс подготовки и преобразования данных, используемый в том числе и для создания изображений, и представлен набор контрольных результатов для набора данных.

1.1 Набор данных Russian Handwritings Tracked

В рамках представленной работы мы создали датасет с данными о написании символов, собрав образцы от 12 писателей. Каждый писатель вносил буквы (строчные и прописные) и цифры. Итого 76 классов образцов данных. Также каждый писатель писал слова из панграммы ”съешь ещё этих мягких французских булок да выпей чаю”, относимые в папку ”extra”. Для каждой пары писатель/символ собрано до 4 образцов, а общее количество образцов в данной версии базы данных составляет 2812. Каждый символ представлен в виде файла трека (траектории), содержащего координаты пера в каждый момент времени при начертании символа. Дополнительно присутствует ряд вспомогательных файлов.

2 Методология

2.1 Сбор и предварительная обработка данных

Образцы почерка были собраны при помощи программы scanner.py на планшете модели XP Pen Deco03 с помощью стилуса. В роли писателей выступали мужчины и женщины в возрасте от 18 до 55 лет. Каждый из 12 писателей выполнил от одного до четырех последовательных сеанса. В каждом сеансе соответствующего писателя просили написать по одному примеру для каждого символа алфавита в фиксированном формате. Набор включал в себя строчные и прописные буквы, цифры, а также слова из панграммы. Программа сбора данных показывает на экране ПК поля строки, и писателям предлагается писать только внутри этих полей. Написание первого и каждого последующего образца символа считается одобренным, если автор принял их как таковые. Программа сбора данных записывает только информацию о координатах X и Y и информацию о времени вдоль хода, без, например, значений уровня давления.

Например, буква ”a” для первого писателя представлена в виде файла **a** со следующим содержимым:

```
"204","259","204","256","202","254","203","251","203","244",
"203","234","205","222","208","219","212","215","215","215",
"222","223","225","233","227","244","229","257","229","263",
"220","270","214","268","207","266","197","254","195","245",
"192","241","195","234","197","231","204","230","227","258",
"227","249","230","241","233","231","237","220","244","209",
"249","206","258","206","263","212","270","218","273","220"
```

При этом надпись автором на экране выглядела следующим образом (рис 1):

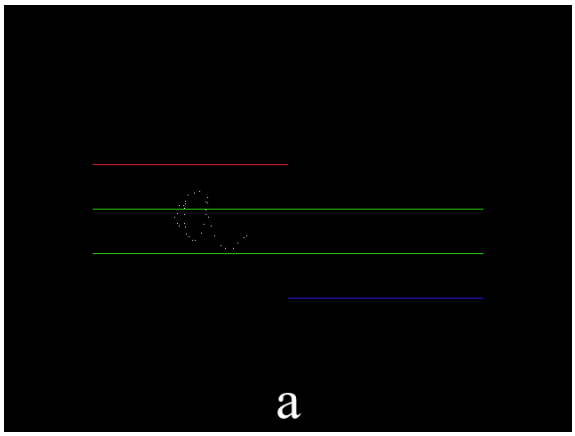

Рис. 1: Символ ”a”при сборе датасета.

Полученные в процессе сбора файлы сохраняются в папку с датасетом в следующей структуре:

```
w_<nw>_<nt>/
    extra/
        <w>
        <w>.png
        <w>_times
        ...
    mnist_like/
        <c>.csv
    <c>
    <c>.png
    <c>_times
    ...
...
convert2mnist.py
scanner.py
```

Где:

- <nw> – номер писателя,
- <nt> – номер попытки,
- <c> – файл с координатами траектории представленного символа,
- <c>.png – файл с изображением символа в том виде, как его видел писатель в процессе написания,
- <c>_times – дополнительный файл с интервалами времени между точками траектории,
- <w> – файл с координатами траектории представленного написания слова из папки extra,
- <w>.png – файл с изображением написания слова в том виде, как его видел писатель в процессе написания,
- <w>_times – дополнительный файл с интервалами времени между точками траектории,
- scanner.py – программа на языке Pyhton, которая применялась в процессе сбора начертаний при создании датасета,
- convert2mnist.py – программа на языке Pyhton, которая использовалась, чтобы трансформировать формат данных траекторий символ в пиксельный (матричный) формат 28×28 данных, подобный датасету MNIST.

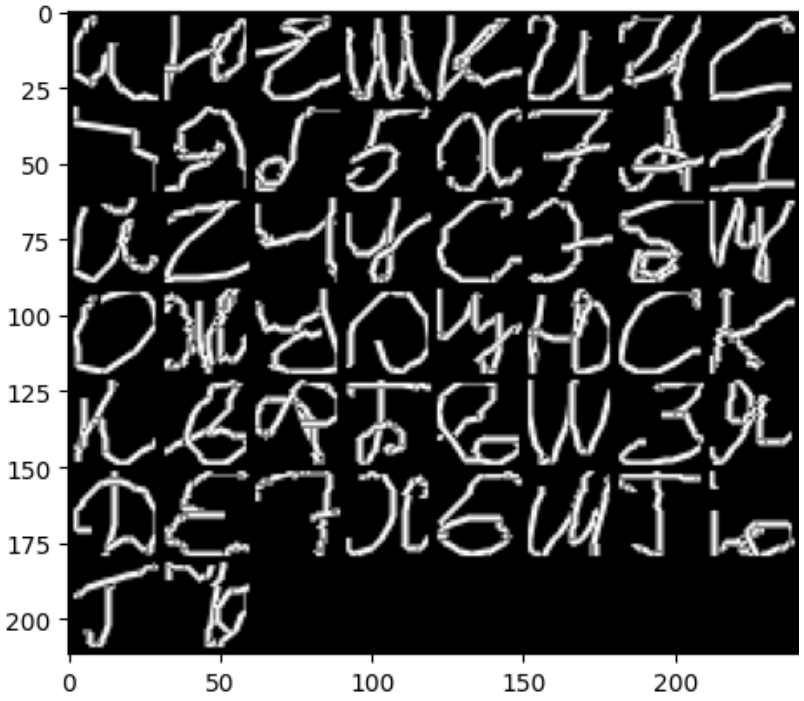

Рис. 2: Пример графических данных Russian Handwritings Tracked.

Пример получившихся изображений для символов *а ю Е ш к И И с 1 я б 5 Х 7 А 1 Й г 4 у С э б щ 0 ж у 0 ц Ю С К к в Ф Б в и 3 я Д Е 7 х 6 Ш Г ю Г ъ* можно увидеть ниже (рис 2).

Распределение меток классов в представленном наборе данных Russian Handwritings Tracked примерно равномерное. На рисунке (рис 3) представлено распределение классов в тренировочной выборке наборе данных .

2.2 Преобразование данных в матричную форму

Представление данных траекторий символов в матричном виде совсем не было основной целью данной работы. Однако для того, чтобы оценить качество собранных данных в сравнении с MNIST-подобными классическим датасетом, траектории были преобразованы в матричный формат 28×28 с нормализованным $[0:1]$ пиксельным представлением символов. Это преобразование выполнялось программой convert2mnist.py, прилагаемой к набору данных. Например, символ ”а” в матричном виде визуализирован на рисунке (рис 4).

2.3 Анализ представленных данных

Несмотря на то что создаваемый набор данных предоставлял другие данные, нежели классические датасеты и предназначен для других форм анализа рукописных символов, однако в рамках представленной работы, был проведен сравнительный анализ с существующими решениями. Для сравнения

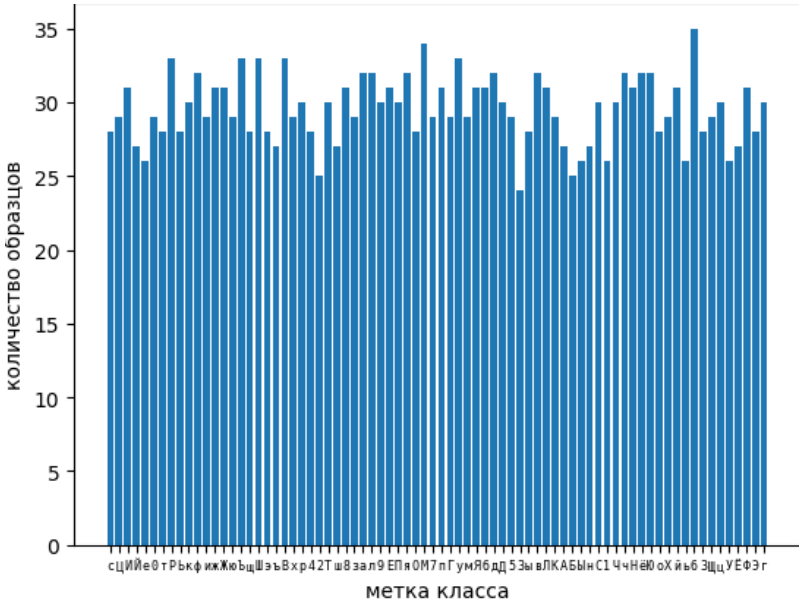

Рис. 3: Распределение классов Russian Handwritings Tracked.

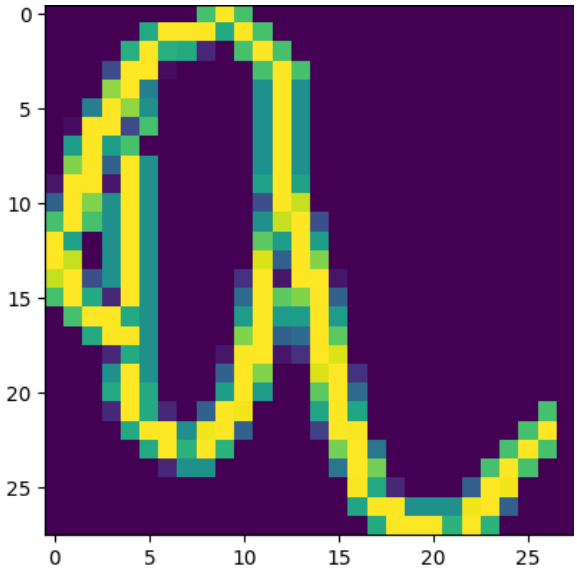

Рис. 4: Матричное представление символа.

с классическими датасетами в качестве базового набора данных был выбран EMNIST [9] с вариантом данных by_class с 62 вариантами символов, а в качестве базовой модели машинного обучения была выбрана LeNET-5 [14]. Структура сверточной сети LeNET-5 представлена в таблице (таб 2) В нашем датасете для анализа было использовано матричное представление символов с 76 классами символов. Для обоих случаев функция потерь использовалась CrossEntropyLoss, метод оптимизации –градиентный спуск с оптимизатором Adam. В процессе обучения были продемонстрированы следующие метрики точности. График точности для датасете EMNIST на рисунке (рис 6). Датасет Russian

Таблица 2: Структура LeNet.

Слой (тип)	Форма выхода	Параметров
Conv2d-1	[-1, 6, 28, 28]	156
Conv2d-2	[-1, 16, 10, 10]	2,416
Linear-3	[-1, 120]	48,120
Linear-4	[-1, 84]	10,164
Linear-5	[-1, 62]	5,270
Total params: 66,126 Trainable params: 66,126 Non-trainable params: 0		

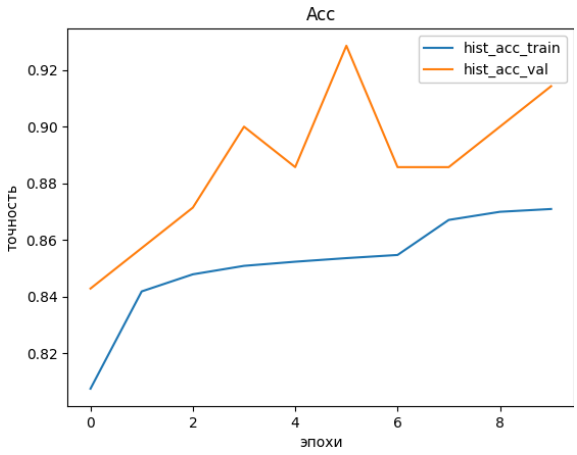

Рис. 5: График метрики точности при обучении на датасете EMNIST.

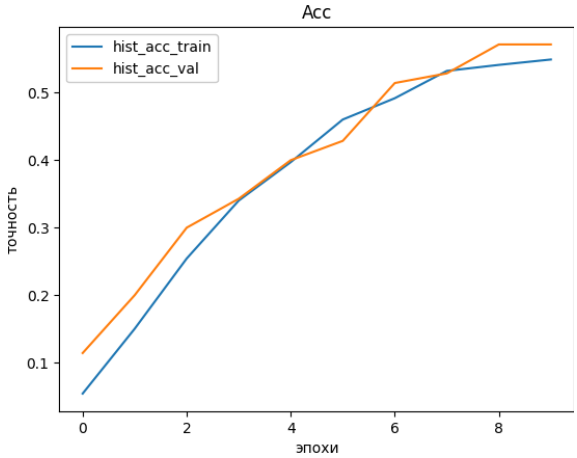

Рис. 6: График метрики точности при обучении на датасете Russian Handwritings Tracked.

Handwritings Tracked показал более скромный результат показанный — 57% точности против 90% у EMNIST. Меньшая точность обусловлена следующими причинами:

- гораздо меньший набор образцов в датасете – 2812 против 697932 у EMNIST
- гораздо большее количество похожих классов – Оо,ШшЩщ,Чч4,Сс,её,ИиЙй,Кк...,

- большее количество классов – 76 против 62 у EMNIST, причем иногда символы больше похожи на изображения других классов, нежели на изображения своего.

Тем не менее показанный уровень точности, если его отнормировать на количество классов вполне сопоставим с показанным EMNIST. Таким образом, можно констатировать, что представленный датасет может быть использован в задах анализ рукописных символов.

3 Результаты

Результаты работы представлены:

- набор данных: `https://data.mendeley.com/datasets/3h6h5d7xg2/2`
- материалы статьи: `https://github.com/d-yacenko/Article10`

В качестве оценки датасета было получено сравнение точности представленного датасета с датасетом EMNIST на базовой модели LeNET-5, которая оказалось несколько хуже эталонного датсета. Однако важно отметить, что Russian Handwritings Tracked обладает набором свойств, отсутствующих у других общеизвестных датасетов, а именно:

- он предоставляет рукописные символы полного русского алфавита,
- он предоставляет информацию о траектории движения пера,
- он дополнительно предоставляет траекторию написания отдельных слов.

4 Выводы

В рамках представленной работы был описан набор данных для обучения моделей ИИ, предназначенных для обработки рукописных символов и текстов на русском языке. Несмотря на более слабые показатели точности на базовой модели, чем у датасета EMNIST, тем не менее точность все же позволяет выполнять классификацию символов в матричном виде с достаточно высокой вероятностью.

Важно отметить, что в датасете имеется информация о движении пера по траектории начертания символов, которая позволяет обогатить данные для моделей обработки рукописного текста. Например, в работе [24] дополнительно была проанализирована динамическая составляющая, что позволило дополнительно увеличить точность классификации на 5%.

Кроме того, описание траектории начертания может позволить извлекать данные об особенности почерка писателя, что может быть использовано для переноса стиля письма, распознавания автора надписи, а также дополнительно улучшать распознавание текста при плохом почерке.

Список литературы

[1] M. Botta, A. Giordana, and L. Saitta, “Learning fuzzy concept definitions,” *[Proceedings 1993] Second IEEE International Conference on Fuzzy Systems*, pp. 18–22 vol.1, 1993.

[2] P. W. Frey and D. J. Slate, “Letter recognition using holland-style adaptive classifiers,” *Machine Learning*, vol. 6, pp. 161–182, 2004.

[3] J. Peltonen, A. Klami, and S. Kaski, “Improved learning of riemannian metrics for exploratory analysis [neural networks 17 (8–9) 1087–1100],” *Neural Networks*, vol. 18, pp. 105–105, 2005.

[4] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, “Online and offline handwritten chinese character recognition: Benchmarking on new databases,” *Pattern Recognition*, vol. 46, no. 1, p. 155–162, 2013.

[5] D.-H. Wang, C.-L. Liu, J.-L. Yu, and X.-D. Zhou, “Casia-olhwdb1: A database of online handwritten chinese characters,” p. 1206–1210, 01 2009.

[6] B. H. Williams, M. Toussaint, and A. J. Storkey, “Extracting motion primitives from natural handwriting data,” in *Artificial Neural Networks – ICANN 2006* (S. Kollias, A. Stafylopatis, W. Duch, and E. Oja, eds.), (Berlin, Heidelberg), p. 634–643, Springer Berlin Heidelberg, 2006.

[7] F. Meier, E. Theodorou, F. Stulp, and S. Schaal, “Movement segmentation using a primitive library,” p. 3407–3412, 09 2011.

[8] T. E. de Campos, B. R. Babu, and M. Varma, “Character recognition in natural images,” in *International Conference on Computer Vision Theory and Applications*, 2009.

[9] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, “EMNIST: an extension of MNIST to handwritten letters,” *CoRR*, vol. abs/1702.05373, 2017.

[10] D. Llorens, F. Prat, A. Marzal, J. M. Vilar, M. J. Castro, J.-C. Amengual, S. Barrachina, A. Castellanos, S. E. Boquera, J. A. Gómez, J. Gorbe-Moya, A. Gordo, V. Palazón, G. Peris, R. Ramos-Garijo, and F. Zamora-Martínez, “The ujipenchars database: a pen-based database of isolated handwritten characters,” in *International Conference on Language Resources and Evaluation*, 2008.

[11] S. Calderara, A. Prati, and R. Cucchiara, “Mixtures of von mises distributions for people trajectory shape analysis,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, pp. 457–471, 2011.

[12] I. M. Guyon, S. R. Gunn, A. Ben-Hur, and G. Dror, “Result analysis of the nips 2003 feature selection challenge,” in *NIPS*, 2004.

[13] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, “Human-level concept learning through probabilistic program induction,” *Science*, vol. 350, no. 6266, p. 1332–1338, 2015.

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, p. 2278–2324, 12 1998.

[15] E. M. Kussul and T. Baidyk, “Improved method of handwritten digit recognition tested on mnist database,” *Image Vis. Comput.*, vol. 22, pp. 971–981, 2004.

[16] L. Xu, A. Krzyzak, and C. Suen, “Methods of combining multiple classifiers and their applications to handwriting recognition,” *IEEE Transactions on Systems Man and Cybernetics Part B (Cybernetics)*, vol. 22, p. 418–435, 06 1992.

[17] F. Alimoglu and E. Alpaydin, “Combining multiple representations and classifiers for pen-based handwritten digit recognition,” *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, vol. 2, pp. 637–640 vol.2, 1997.

[18] E. Tang, P. Suganthan, X. Yao, and K. Qin, “Linear dimensionality reduction using relevance weighted lda,” *Pattern Recognition*, vol. 38, p. 485–493, 04 2005.

[19] Y. Hong, Q. Li, J. Jiang, and Z. Tu, “Learning a mixture of sparse distance metrics for classification and dimensionality reduction,” *2011 International Conference on Computer Vision*, pp. 906–913, 2011.

[20] M. Thoma, “The hasyv2 dataset,” *ArXiv*, vol. abs/1701.08380, 2017.

[21] M. Karki, Q. Liu, R. Dibiano, S. Basu, and S. Mukhopadhyay, “Pixel-level reconstruction and classification for noisy handwritten bangla characters,” p. 511–516, 08 2018.

[22] Q. Liu, E. Collier, and S. Mukhopadhyay, “Pcgan-char: Progressively trained classifier generative adversarial networks for classification of noisy handwritten bangla characters,” in *International Conference on Asian Digital Libraries*, 2019.

[23] Dmitry Iatsenko and Konstantin Smirnov, “Russian handwritings tracked,” *Mendeley*, 2023.

[24] I. Saenko, O. Lauta, and D. Iatsenko, “The use of dynamic characteristics in handwriting recognition tasks,” in *2023 International Ural Conference on Electrical Power Engineering (UralCon)*, pp. 609–614, 2023.