

Гарантированное покрытие типов аугментации на наборе Digits

М. Д. Кашпиров

Аннотация - Проверяется гипотеза о преимуществе гарантированного покрытия типов аугментации перед случайной смесью при одинаковом объёме обучающих данных. Сбалансированная многотиповая аугментация СБМТА-5 формирует для каждого исходного изображения по одной копии из пяти семейств преобразований; RandomMix-5 независимо выбирает семейство для каждой из пяти копий. Сравнение выполнено на наборе Digits из 1797 изображений размером 8×8 для RBF-SVM и логистической регрессии. Использованы двадцать повторных стратифицированных разбиений 70/30, семь режимов искажений представленных семейств и четыре режима семейств, отсутствующих в обучающей политике. Стохастические тестовые искажения усреднялись по десяти реализациям внутри разбиения. Парные различия оценивались скорректированным тестом Надо–Бенжиро с поправкой Холма. Преимущество СБМТА-5 перед RandomMix-5 не подтверждено ни для одного классификатора: гипотеза о дополнительном выигрыше от покрытия в исследованных условиях не получила поддержки. Этот результат не доказывает эквивалентность методов. По сравнению с обучением на чистых данных обе многотиповые схемы повышают среднюю и минимальную точность на проверенном профиле искажений. Для RBF-SVM средняя точность СБМТА-5 составляет 89,96 % против 84,19 % у исходной модели, однако при размывании она снижается с 88,26 до 82,80 %. Результаты относятся к компактному представлению Digits и выбранным синтетическим преобразованиям; перенос на изображения большего разрешения и реальные дефекты не установлен. Код и исходные таблицы приведены в дополнительных материалах.

Ключевые слова — аугментация данных, покрытие преобразований, классификация цифр, метод опорных векторов, логистическая регрессия, устойчивость к искажениям.

I. ВВЕДЕНИЕ

Аугментация часто повышает точность классификации искажённых изображений, однако её эффект зависит от состава преобразований и условий проверки [2], [3]. Одна из возможных политик состоит в том, чтобы применить к каждому обучающему объекту все выбранные типы преобразований. Такая политика гарантирует покрытие типов, но сама по себе не устанавливает преимущество по качеству модели. В этой работе проверяется, даёт ли полное покрытие каждого объекта дополнительный выигрыш по сравнению со случайным выбором из тех же типов при одинаковом числе обучающих примеров.

AutoAugment и RandAugment предлагают способы выбора политик преобразований [6], [7]. AugMix смешивает цепочки операций для повышения

устойчивости к распространённым искажениям [8]. Mixup, Cutout, Random Erasing и Patch Gaussian изменяют изображения или отдельные области [9]–[12], а IPMix сочетает воздействия на уровне изображения, фрагментов и пикселей [13]. Эти подходы изучались преимущественно для глубоких сетей. Их результаты не позволяют заранее заключить, что обязательное покрытие типов будет полезно для классических классификаторов на компактных данных.

Основной вопрос исследования - превосходит ли сбалансированная многотиповая аугментация СБМТА-5 равнобюджетную случайную смесь RandomMix-5. Дополнительные сравнения с исходными данными, их дублированием и односемейной аугментацией позволяют оценить вклад состава преобразований. Анализ выполнен для двух классификаторов, отдельно для семейств, представленных в обучающей политике, и семейств, отсутствующих в ней.

Работа представляет контролируемое сравнение двух способов распределения одного набора преобразований на Digits. Центральный результат отрицателен: гипотеза о дополнительном преимуществе покрытия не подтверждена. Улучшение относительно чистого обучения наблюдается и у случайной смеси, поэтому его нельзя приписывать гарантии покрытия. Область выводов ограничена изображениями 8×8 и заданным профилем синтетических искажений. Анализ по классам и отдельным режимам показывает, какие ухудшения скрывает средняя точность.

II. СВЯЗАННЫЕ ИССЛЕДОВАНИЯ

Тестирование на распространённых искажениях выявляет ограничения, которые не видны по точности на исходном распределении [3]. Работа о shortcut learning связывает часть таких ограничений с использованием нестабильных признаков [4]. Mintun и соавторы исследовали зависимость эффекта аугментации от близости обучающих и тестовых искажений [5]. Поэтому семейства, отсутствующие в обучающей политике, здесь рассматриваются отдельно. Название «новые искажения» означает только отсутствие соответствующего семейства в этой политике и не подразумевает произвольный будущий сдвиг распределения.

Для линейных моделей и SVM повреждение признаков также используется при обучении. В [14] исследовано обучение SVM с усреднением потерь по распределению повреждений. В рассматриваемом эксперименте преобразованные копии формируются явно, без изменения алгоритмов обучения. СБМТА-5 задаёт

правило распределения пяти известных семейств между объектами. Эмпирический вопрос состоит в полезности этого правила; новизна элементарных преобразований не заявляется.

III. СБАЛАНСИРОВАННАЯ МНОГОТИПОВАЯ АУГМЕНТАЦИЯ

А. Постановка задачи

Пусть $D = \{(x_i, y_i)\}_{i=1}^N$ - обучающая выборка, где $x_i \in [0, 1]^{64}$ представляет изображение 8×8 , а $y_i \in \{0, \dots, 9\}$. Используются пять семейств $T = \{T_G, T_M, T_S, T_I, T_C\}$: гауссовский шум, маскирование пикселей, сдвиг, импульсный шум и сочетание шума с маскированием. Преобразованным копиям присваивается исходная метка. Её семантическое сохранение предполагается политикой, но не проверялось независимой разметкой.

В. Гарантия покрытия

СБМТА-5 добавляет по одной копии каждого объекта из каждого семейства. При N исходных объектах обучающая матрица содержит $6N$ строк. Объединение в (1) понимается как соединение списков размеченных примеров с сохранением повторов:

$$D_{CB} = D \cup T_1(D) \cup \dots \cup T_5(D). \quad (1)$$

Коэффициент покрытия c_i равен доле семейств, хотя бы один раз применённых к объекту x_i . Для СБМТА-5 $c_i = 1$. В RandomMix-5 также создаются пять копий, но их семейства выбираются независимо и равновероятно из того же набора. Ожидаемое покрытие составляет

$$E[c_i] = 1 - (1 - \frac{1}{5})^5 = 0,67232. \quad (2)$$

Следовательно, RandomMix-5 в среднем не использует около трети семейств для отдельного объекта. На уровне всей выборки ожидаемые частоты семейств одинаковы у двух схем. Различие иллюстрирует рис. 1. Полное покрытие является свойством процедуры формирования данных; его влияние на точность проверяется экспериментально.



Рис. 1. Сбалансированное и случайное покрытие пяти семейств преобразований.

ТАБЛИЦА I. Число различных семейств в RandomMix-5

Число семейств	Вероятность %
1	0,16
2	9,60
3	48,00
4	38,40
5	3,84

Вероятность полного покрытия в RandomMix-5 равна $5!/5^5 = 3,84\%$. Наиболее вероятны три различных семейства (48,00 %), а ожидаемое число семейств составляет 3,3616. Эти значения получены из модели

независимых выборов и не являются оценками качества классификации. Число различных семейств также не характеризует степень изменения изображения или семантическую различимость полученных копий.

С. Параметры преобразований

Для T_G стандартное отклонение шума независимо для каждого изображения выбирается из $U[0,06; 0,22]$. Для T_M вероятность обнуления пикселя выбирается из $U[0,08; 0,25]$, для T_I вероятность замены пикселя значением 0 или 1 - из $U[0,04; 0,14]$. T_S выбирает одно из четырёх направлений сдвига на один пиксель; освободившаяся область заполняется нулями, вышедшие за границу значения отбрасываются. T_C добавляет гауссовский шум с $\sigma = 0,12$, затем маскирует пиксели с вероятностью 0,08. После добавления шума значения ограничиваются интервалом $[0, 1]$. Диапазоны фиксированы для основного эксперимента.

Фиксированное покрытие не устраняет случайность преобразований. Для каждого разбиения используется одна обучающая реализация аугментации на метод. Она общая для двух классификаторов; десятикратное повторение относится только к стохастическим искажениям тестовых данных.

IV. ЭКСПЕРИМЕНТАЛЬНЫЙ ПРОТОКОЛ

А. Данные, модели и разбиения

Использован `load_digits` библиотеки `scikit-learn`: 1797 изображений десяти классов размером 8×8 [20]. Это тестовая часть Optical Recognition of Handwritten Digits из UCI [17], а не уменьшенная версия MNIST [22]. Представление UCI получено из нормализованных бинарных изображений 32×32 : в неперекрывающихся блоках 4×4 подсчитаны активные пиксели. Таким образом, 64 признака со значениями 0–16 описывают плотность штрихов в блоках. Перед моделированием они делились на 16. Компактность набора позволяет провести повторные сравнения, однако существенно ограничивает внешнюю применимость результатов.

Выполнено двадцать повторных стратифицированных разбиений 70/30 с `seed` 41–60. В каждом разбиении обучающая часть содержит 1257 объектов, тестовая - 540. Аугментация применяется только к обучающей части. Для всех методов используются одинаковые разбиения и реализации тестовых искажений. Основная модель — RBF-SVM с $C=10$ и $\gamma=\text{scale}$ [1]; вторая модель - многоклассовая логистическая регрессия с $C=1$, `solver=lbfgs` и `max_iter=3000`. Значение $\gamma=\text{scale}$ определяется по дисперсии соответствующей обучающей матрицы и поэтому зависит от политики аугментации. Реализация использует `scikit-learn` [18] и `SciPy` [19].

В. Равнобюджетные базовые варианты

ТАБЛИЦА II. ОБЪЁМ ВЫБОРКИ И ПОКРЫТИЕ ТИПОВ ПРЕОБРАЗОВАНИЙ

Метод	Объём	Покрытие
Clean	N	—
Clean×6	6N	0
Gaussian×5	6N	1/5
RandomMix-5	6N	$E[c_i]=0,67232$
СБМТА-5	6N	1

Clean×6 проверяет, достаточно ли увеличения числа строк без новых изображений. Дублирование при фиксированном C может менять соотношение регуляризации и потерь, поэтому его результат нельзя трактовать как чистую оценку размера выборки. Главным контролем покрытия служит RandomMix-5 с теми же семействами и распределениями интенсивностей. Gaussian×5 проверяет односемейную аугментацию. Все варианты, кроме Clean, содержат 7542 строки обучающих данных на разбиение. СБМТА-5 балансирует типы преобразований внутри объекта; исходные пропорции классов сохраняются, специальная балансировка классов не выполняется.

С. Тестовые искажения и метрики

Профиль содержит семь режимов из представленных семейств и четыре режима отсутствующих семейств: поворот на $+10^\circ$ или -10° , гауссовское размытие $\sigma=0,8$ пикселя, уменьшение контраста с коэффициентом 0,55 и дилатацию с окном 2×2 . Отнесение к представленным семействам не означает совпадения интенсивности: например, тестовый шум $\sigma=0,30$ выходит за обучающий диапазон. Для восьми стохастических режимов генерируется по десять реализаций на разбиение; чистые данные и три детерминированных искажения оцениваются по одному разу. На пару «модель–метод» приходится 1680 записей оценки, всего — 16 800. Это число вычислительных оценок, а не независимых наблюдений.

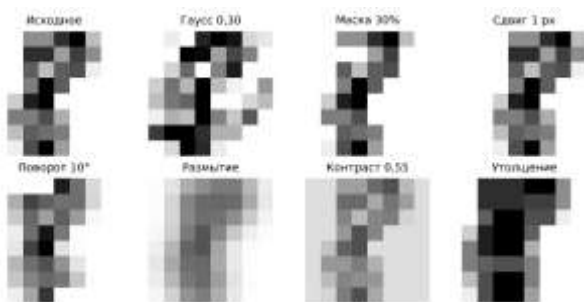


Рис. 2. Примеры представленных и непредставленных при обучении искажений.

ТАБЛИЦА III. ПРОФИЛЬ ТЕСТОВЫХ ИСКАЖЕНИЙ

Семейство	Группа	Уровень
Гауссовский шум	Предст.	$\sigma=0,15; 0,30$
Маскирование	Предст.	15; 30 %
Импульсный шум	Предст.	15 %
Сдвиг	Предст.	1 пиксель
Комбинация	Предст.	$\sigma=0,18$; маска 12 %
Поворот	Отсутст.	$+10^\circ$ или -10°
Размытие	Отсутст.	$\sigma=0,8$ пикселя
Контраст	Отсутст.	коэффициент 0,55
Дилатация	Отсутст.	окно 2×2

Десять реализаций каждого стохастического режима сначала усредняются внутри разбиения. Средняя точность A_{mean} затем вычисляется с равными весами режимов, а A_{worst} — как минимум их усреднённых точностей. Эти показатели определяются отдельно для семи представленных, четырёх отсутствующих и всех одиннадцати режимов. В таблицах приведены средние полученных показателей по двадцати разбиениям. A_{worst} характеризует худший из проверенных режимов и не является гарантией устойчивости к любому возмущению. Macro-F1 сохранён в исходных таблицах; статистическая единица анализа — разбиение.

Из-за перекрытия повторных разбиений применяется двусторонний corrected resampled t-test Надо–Бенжю [15]. Для двадцати парных разностей стандартная ошибка равна квадратному корню из произведения их выборочной дисперсии на $(1/20+540/1257)$; число степеней свободы равно 19. Поправка Холма применяется к шести показателям робастности отдельно для каждой пары «классификатор–базовый метод» [16]. Приведённые 95%-ные интервалы учитывают коррекцию повторных разбиений, но не являются одновременными интервалами для всего семейства сравнений. Коррекция приближённо учитывает зависимость оценок и не превращает разбиения в независимые наборы данных.

V. РЕЗУЛЬТАТЫ

А. RBF SVM при одинаковом объёме обучающих данных

ТАБЛИЦА IV. СРЕДНИЕ ПОКАЗАТЕЛИ RBF-SVM, %

Метрика	Clean	Clean ×6	Gauss. ×5	Random Mix-5	СБМТА-5
Чистые	98,92	98,91	98,90	98,90	98,93
Предст. ср.	85,43	85,43	86,06	92,50	92,52
Предст. мин.	55,37	55,38	56,04	85,83	85,71
Отсутст. ср.	82,01	81,99	85,23	85,91	85,48
Отсутст. мин.	62,10	62,03	64,65	77,66	77,98
Все ср.	84,19	84,18	85,76	90,10	89,96
Все мин.	55,37	55,38	56,04	77,66	77,98

Предст. — семейства обучающей политики; отсутст. — отсутствующие семейства. «Ср.» и «мин.» обозначают среднюю и минимальную точность по режимам внутри разбиения.

Точность СБМТА-5 на чистых данных составляет 98,93 %, у Clean — 98,92 %. Это близкие точечные оценки, а не доказательство отсутствия потерь. По сравнению с

Clean средняя точность на представленных семействах увеличивается на 7,08 п.п., минимум — на 30,34 п.п. На отсутствующих семействах прирост составляет 3,46 и 15,88 п.п. Средняя точность по всем искажениям возрастает с 84,19 до 89,96 %, минимум — с 55,37 до 77,98 %. Для последних двух показателей разность положительна во всех двадцати разбиениях; $p_{Holm} < 0,001$.

Clean×6 практически совпадает с Clean для RBF-SVM. Gaussian×5 даёт 85,76 % в среднем по всем режимам, СБМТА-5 — 89,96 %; разность минимумов составляет 21,94 п.п. В пределах этого профиля односемейная аугментация уступает многотиповой. Однако сравнение с Gaussian×5 одновременно меняет состав преобразований и их распределение, поэтому оно не доказывает пользу гарантированного покрытия.

RandomMix-5 достигает 90,10 % в среднем по всем искажениям. Разность СБМТА-5–RandomMix-5, рассчитанная до округления, равна −0,15 п.п. с 95%-ным ДИ [−0,87; +0,58]. Для минимума разность составляет +0,32 п.п., ДИ [−4,55; +5,20]. Для обоих показателей $p_{Holm} = 1,000$. По всем шести показателям робастности преимущество СБМТА-5 не подтверждено. Следовательно, улучшение относительно Clean не служит доказательством дополнительного эффекта покрытия.

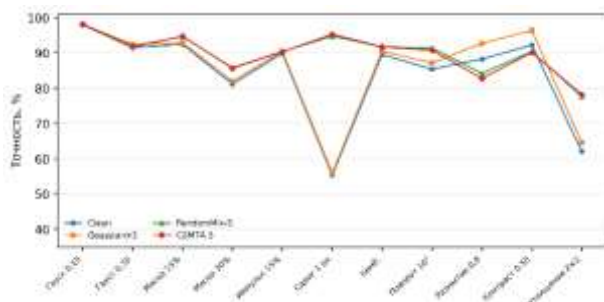


Рис. 3. Точность RBF-SVM в отдельных режимах.

По отдельным режимам многотиповые схемы повышают точность при сдвиге, маскировании, повороте и дилатации. При размытии СБМТА-5 даёт 82,80 % против 88,26 % у Clean и 92,75 % у Gaussian×5. При снижении контраста соответствующие значения равны 90,16, 92,27 и 96,37 %. Поэтому средний выигрыш сопровождается ухудшением в части отсутствующих семейств. Точки на рис. 3 усреднены по двадцати разбиениям, для стохастических режимов — также по десяти реализациям внутри разбиения.

В. Результаты логистической регрессии

Таблица V. Средние показатели логистической регрессии, %

Метрика	Clean	Clean ×6	Gauss. ×5	Random Mix-5	СБМТА-5
Чистые	96,59	96,79	96,69	95,59	95,57
Предст. ср.	79,86	77,79	79,28	81,59	81,76
Предст. мин.	42,80	42,16	43,19	63,56	63,84
Отсутст. ср.	77,44	73,75	78,33	83,96	84,56
Отсутст. мин.	63,14	59,09	63,34	68,89	70,67
Все ср.	78,98	76,32	78,93	82,45	82,78
Все мин.	42,80	42,16	43,19	63,46	63,84

Предст. - семейства обучающей политики; отсутст. - отсутствующие семейства. «Ср.» и «мин.» обозначают

среднюю и минимальную точность по режимам внутри разбиения.

Для логистической регрессии СБМТА-5 снижает точность на чистых данных с 96,59 до 95,57 %. Средняя точность по всем искажениям возрастает с 78,98 до 82,78 %, минимум - с 42,80 до 63,84 %. Для среднего и минимума разность с Clean положительна на всех двадцати разбиениях и значима после поправки Холма ($p < 0,001$). Таким образом, вторая модель также выигрывает от многотипового обучения на выбранном профиле, но теряет около 1,02 п.п. на исходных изображениях.

В сравнении с RandomMix-5 СБМТА-5 даёт +0,60 п.п. по среднему и +1,78 п.п. по минимуму на отсутствующих семействах. Однако интервалы [−0,46; +1,66] и [−1,26; +4,82] п.п. включают ноль. Для среднего по всем искажениям разность составляет +0,33 п.п., ДИ [−0,26; +0,91]. По всем шести показателям $p_{Holm} = 1,000$. Преимущество покрытия не подтверждается и для логистической регрессии.

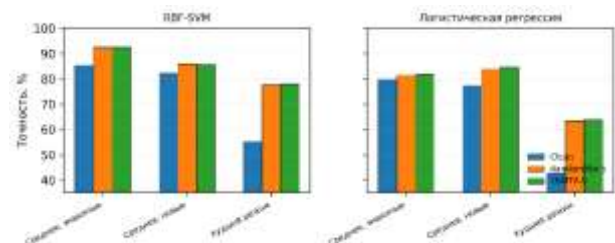


Рис. 4. Агрегированная устойчивость двух классификаторов.

С. Парные различия и их неопределённость

Таблица VI. Парные разности СБМТА-5 с базовыми методами

Мод.	База и метрика	Δ п.п.	95% ДИ п.п.	p_{Holm}
SVM	Clean Все ср.	+5,77	[+4,90; +6,64]	<0,001
SVM	Clean Все мин.	+22,61	[+18,21; +27,02]	<0,001
SVM	RM-5 Все ср.	−0,15	[−0,87; +0,58]	1,000
SVM	RM-5 Все мин.	+0,32	[−4,55; +5,20]	1,000
ЛР	RM-5 Отс. ср.	+0,60	[−0,46; +1,66]	1,000
ЛР	RM-5 Отс. мин.	+1,78	[−1,26; +4,82]	1,000
ЛР	RM-5 Все ср.	+0,33	[−0,26; +0,91]	1,000

Δ=СБМТА-5–база; RM-5 - RandomMix-5, ЛР - логистическая регрессия. Интервалы учитывают коррекцию перекрывающихся разбиений.

Таблица VI показывает разности СБМТА-5 с базовыми методами. Сравнения с Clean подтверждают улучшение среднего и минимума на выбранном профиле, а сравнения с RandomMix-5 не подтверждают преимущество покрытия. Незначимый результат не доказывает равенство методов [21]. Особенно широк интервал разности минимумов для RBF-SVM: данные совместимы как с ухудшением, так и с улучшением на несколько процентных пунктов. Граница практически существенного эффекта заранее не задавалась; тест эквивалентности не проводился.

D. Анализ по классам

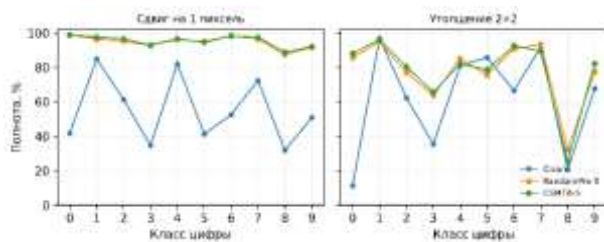


Рис. 5. Полнота отдельных классов RBF-SVM при сдвиге и утолщении.

При сдвиге средняя по разбиениям полнота отдельных классов исходной SVM составляет 31,79–84,83 %, размах - 53,04 п.п. Для СБМТА-5 диапазон равен 88,65–98,78 %, размах - 10,12 п.п. При дилатации минимальная средняя полнота повышается с 11,30 до 25,29 %, размах сокращается с 85,52 до 70,71 п.п. RandomMix-5 имеет близкий общий профиль. Эти покласовые значения описательны: отдельные проверки значимости для всех классов не выполнялись.

При размытии минимальная средняя полнота класса для СБМТА-5 составляет 14,91 %, у Clean - 61,27 %. Этот результат согласуется со снижением общей точности в данном режиме. Повышение агрегированных показателей не обеспечивает улучшения для каждого класса. Гауссовский шум и размытие нельзя считать взаимозаменяемыми только потому, что оба изменяют локальные значения яркости.

E. Чувствительность к интенсивности

При повышении σ гауссовского шума с 0,15 до 0,30 точность исходной SVM снижается на 6,48 п.п., СБМТА-5 - на 6,35 п.п. При повышении вероятности маскирования с 15 до 30 % снижение составляет 11,30 и 9,03 п.п. соответственно. Эти описательные сравнения относятся к двум выбранным уровням; они не устанавливают общего преимущества покрытия при росте интенсивности.

Полноценная зависимость точности от интенсивности не исследовалась. Для её оценки нужен заранее заданный диапазон уровней и одинаковое правило агрегирования. При сопоставлении разрешений параметры геометрических и пространственных преобразований следует нормировать по размеру изображения, а также проверять сохранение распознаваемого класса.

F. Вычислительная стоимость

ТАБЛИЦА VII. Средняя стоимость обучения

Модель метод	Строк	Обучение с	Опорных векторов
SVM Clean	1257	0,028	531
SVM RM-5	7542	0,764	2624
SVM СБМТА-5	7542	0,783	2677
ЛР Clean	1257	0,053	—
ЛР RM-5	7542	0,420	—
ЛР СБМТА-5	7542	0,414	—

В использованной среде обучение RBF-SVM занимает в среднем 0,028 с для Clean, 0,764 с для RandomMix-5 и 0,783 с для СБМТА-5; среднее число опорных векторов составляет 531, 2624 и 2677. Для логистической регрессии время увеличивается с 0,053 до 0,414 с при

переходе от Clean к СБМТА-5. Абсолютные значения зависят от среды и не являются внешним сравнением производительности. Тестовая аугментация не применяется, но рост числа опорных векторов может увеличить стоимость предсказания SVM.

VI. ОБСУЖДЕНИЕ

Главная проверяемая гипотеза о дополнительном выигрыше гарантированного покрытия не получила поддержки. СБМТА-5 и RandomMix-5 дают близкие точечные оценки, а скорректированный анализ не выявляет различий по шести показателям робастности ни для одной модели. В исследованной конфигурации нет эмпирического основания выбирать СБМТА-5 ради повышения точности относительно RandomMix-5. При этом отсутствие подтверждённого преимущества не позволяет утверждать, что покрытие бесполезно при любых объёмах данных и условиях.

Улучшение относительно Clean получено у обеих многотиповых схем. Для SVM дублирование не воспроизводит этот выигрыш, а односемейная аугментация даёт менее широкий профиль. Результат поддерживает использование нескольких выбранных семейств в данном эксперименте, но не обязательное покрытие каждого объекта. Кроме того, ухудшение при размытии и снижении контраста показывает, что такой выигрыш нельзя распространять на все искажения.

Правило СБМТА-5 позволяет непосредственно проверить наличие всех пяти типов копий для каждого объекта. Это организационное свойство обучающей выборки, а не метрика качества модели. Семейства в СБМТА-5 назначаются фиксированно, однако интенсивности шума, маски и направления сдвигов остаются случайными. Обе схемы воспроизводятся при фиксированных генераторах, версиях библиотек и параметрах. Поэтому воспроизводимость и возможность аудита не заменяют неподтверждённый выигрыш по точности.

VII. ОГРАНИЧЕНИЯ ИССЛЕДОВАНИЯ И ВОСПРОИЗВОДИМОСТЬ

Разрешение 8×8 существенно меняет смысл выбранных воздействий. Сдвиг на один пиксель равен 12,5 % ширины или высоты изображения и в используемой реализации может удалять крайние фрагменты штриха. Поворот на $\pm 10^\circ$ выполняется с линейной интерполяцией, сохранением размера и нулевым заполнением за границей; его эффект включает перераспределение яркости и обрезание. Размытие с $\sigma=0,8$ пикселя имеет масштаб 10 % стороны изображения, а окно дилатации 2×2 охватывает четверть её линейного размера. На такой сетке разные операции могут сходным образом менять несколько соседних элементов. Их перцептивная различимость отдельно не измерялась.

Преобразования выполняются над уже агрегированными признаками 8×8 , а не над исходными сканами. Поэтому наблюдаемый перенос между отсутствующими семействами характеризует это представление и не устанавливает устойчивость к

дефектам получения реальных изображений. Исходная метка сохраняется программно; возможная потеря читаемости после сильного преобразования не проверялась независимыми экспертами.

Набор содержит примерно одинаковое число объектов каждого класса; эффекты сильного дисбаланса не исследованы. Повторные стратифицированные разбиения используют одну и ту же коллекцию и не обеспечивают независимой проверки на другом источнике данных. В протоколе отсутствует разделение по авторам написания, поэтому он не устанавливает перенос на новый почерк.

В среднем эффективное γ для SVM равно 0,110514 у Clean, 0,114092 у RandomMix-5 и 0,114129 у СБМТА-5. Сравнение характеризует процедуры с γ =scale и не полностью отделяет действие преобразований от изменения ширины ядра. Нужна проверка с фиксированным γ . Кроме того, одна обучающая реализация аугментации на разбиение не позволяет отдельно оценить чувствительность к случайному составу обучающих копий.

Дополнительные материалы содержат код, 16 800 исходных записей ассигасу и macro-F1, результаты агрегации, парные статистические сравнения, 7200 поклассовых оценок полноты, данные о стоимости и эффективном γ , параметры среды и контрольные суммы. Основная часть включает 200 обучений; ещё 60 повторных обучений RBF-SVM выполнены для поклассового анализа. Число 9 072 000 соответствует объёму предсказаний, использованных для основных агрегированных оценок; сами массивы индивидуальных предсказаний в архиве не сохранены.

А. Условия дальнейшей проверки гипотезы

Влияние покрытия при меньшем числе исходных объектов, выраженном дисбалансе классов и целенаправленных атаках в этой работе не проверялось. Эти условия следует рассматривать как отдельные гипотезы. Для проверки зависимости от объёма данных нужны несколько заранее заданных размеров обучающей части при неизменном тестовом наборе, а для дисбаланса — контролируемые пропорции классов и поклассовые показатели. Для атак необходимы заданная модель угроз и самостоятельный протокол оценки.

Проверка внешней применимости требует независимого эксперимента на MNIST [22] или других изображениях большего разрешения и на реальных сканах. Интенсивности сдвига и размытия следует сопоставлять в относительных единицах, учитывать способы интерполяции и обработки границ. Для проверки переноса на новый почерк необходимо разделение по авторам написания. До такой проверки вывод о преимуществе или отсутствии преимущества покрытия относится только к использованной конфигурации Digits.

VIII. ЗАКЛЮЧЕНИЕ

На наборе Digits гипотеза о преимуществе СБМТА-5 перед равнобюджетной случайной смесью не подтверждена ни для RBF-SVM, ни для логистической

регрессии. Отсутствие значимых различий не доказывает эквивалентность методов. Обе многотиповые схемы повышают среднюю и минимальную точность по выбранным искажениям относительно обучения на чистых данных; этот эффект нельзя приписывать гарантии покрытия. Для RBF-SVM СБМТА-5 даёт прирост среднего на 5,77 п.п. и минимума на 22,61 п.п., но снижает точность при размытии. Полное покрытие остаётся проверяемым свойством формирования выборки, без установленного дополнительного выигрыша по качеству. Результаты ограничены изображениями 8×8 и заданными синтетическими операциями; их перенос на другие разрешения и реальные условия требует независимой проверки.

БИБЛИОГРАФИЯ

- [1] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273-297, 1995, doi: 10.1007/BF00994018.
- [2] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, art. no. 60, 2019, doi: 10.1186/s40537-019-0197-0.
- [3] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proc. ICLR*, 2019. Available: <https://openreview.net/forum?id=HJz6tiCqYm>
- [4] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, pp. 665-673, 2020, doi: 10.1038/s4256-020-00257-z.
- [5] E. Mintun, A. Kirillov, and S. Xie, "On interaction between augmentations and corruptions in natural corruption robustness," in *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 3571-3583, 2021. Available: <https://proceedings.neurips.cc/paper/2021/hash/1d49780520898fe37f0cd6b41c5311bf-Abstract.html>
- [6] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. CVPR*, pp. 113-123, 2019, doi: 10.1109/CVPR.2019.00020.
- [7] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 18613-18624, 2020. Available: <https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html>
- [8] D. Hendrycks et al., "AugMix: A simple data processing method to improve robustness and uncertainty," in *Proc. ICLR*, 2020. Available: <https://openreview.net/forum?id=S1gmrxHFvB>
- [9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018. Available: <https://openreview.net/forum?id=r1Ddp1-Rb>
- [10] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with Cutout," *arXiv:1708.04552*, 2017, doi: 10.48550/arXiv.1708.04552.
- [11] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 13001-13008, 2020, doi: 10.1609/aaai.v34i07.7000.
- [12] R. G. Lopes, D. Yin, B. Poole, J. Gilmer, and E. D. Cubuk, "Improving robustness without sacrificing accuracy with Patch Gaussian augmentation," *arXiv:1906.02611*, 2019, doi: 10.48550/arXiv.1906.02611.
- [13] Z. Huang et al., "IPMix: Label-preserving data augmentation method for training robust classifiers," in *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 63660-63673, 2023. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/c917d8b9e01427f3184d80ade22f4d1f-Abstract-Conference.html
- [14] N. Chen, J. Zhu, J. Chen, and B. Zhang, "Dropout training for support vector machines," in *Proc. AAAI Conf. Artif. Intell.*, vol. 28, no. 1, pp. 1752-1759, 2014, doi: 10.1609/aaai.v28i1.8965.
- [15] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Mach. Learn.*, vol. 52, no. 3, pp. 239-281, 2003, doi: 10.1023/A:1024068626366.
- [16] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.*, vol. 6, no. 2, pp. 65-70, 1979. Available: <https://www.jstor.org/stable/4615733>
- [17] E. Alpaydin and C. Kaynak, "Optical Recognition of Handwritten Digits," *13001 Mach. Learn. Repository*, 1998, doi: 10.24432/C50P49.

- [18] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825-2830, 2011. Available: <https://jmlr.org/papers/v12/pedregosa11a.html>
- [19] P. Virtanen et al., “SciPy 1.0: Fundamental algorithms for scientific computing in Python,” *Nat. Methods*, vol. 17, pp. 261-272, 2020, doi: 10.1038/s41592-019-0686-2.
- [20] Scikit-learn developers, “load_digits,” scikit-learn documentation. Available: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html (accessed Oct. 11, 2026).
- [21] D. Lakens, “Equivalence tests: A practical primer for t tests, correlations, and meta-analyses,” *Soc. Psychol. Personal. Sci.*, vol. 8, no. 4, pp. 355–362, 2017, doi: 10.1177/1948550617697177.
- [22] Y. LeCun, C. Cortes, and C. J. C. Burges, “The MNIST database of handwritten digits.” Available: <https://yann.lecun.org/exdb/mnist/> (accessed Oct. 11, 2026).

Guaranteed Augmentation Coverage on the Digits Dataset

Mikhail D. Kashpirev

Abstract - This study tests whether guaranteed transformation coverage improves classification accuracy over random mixing at the same training-data budget. Coverage-balanced multi-perturbation augmentation (CB-MPA-5) adds one copy from each of five transformation families to every training image. RandomMix-5 independently selects a family for each of five copies. Experiments use 1,797 Digits images of size 8×8, RBF-SVM, logistic regression, and twenty repeated stratified 70/30 holdouts. Seven corruption regimes belong to training families; four belong to families absent from the training policy. Ten realizations of each stochastic test corruption are averaged within each holdout. Paired comparisons use the Nadeau–Bengio corrected resampled t-test with Holm adjustment. The hypothesis of an additional benefit from guaranteed coverage is not supported for either classifier. This does not establish equivalence. Both multi-perturbation policies improve aggregate corrupted accuracy over clean training. For RBF-SVM, CB-MPA-5 raises mean accuracy from 84.19% to 89.96%, but reduces accuracy under blur from 88.26% to 82.80%. Coverage describes sample construction rather than demonstrated model superiority. Findings are restricted to this compact dataset and the specified synthetic corruptions; transfer to higher-resolution images or real acquisition defects remains untested. Code and aggregate source results accompany the article.

Keywords - data augmentation, transformation coverage, digit classification, support vector machine, logistic regression, corruption robustness.

REFERENCES

- [1] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.
- [2] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, art. no. 60, 2019, doi: 10.1186/s40537-019-0197-0.
- [3] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proc. ICLR*, 2019. Available: <https://openreview.net/forum?id=HJz6tiCqYm>
- [4] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, pp. 665–673, 2020, doi: 10.1038/s42256-020-00257-z.
- [5] E. Mintun, A. Kirillov, and S. Xie, "On interaction between augmentations and corruptions in natural corruption robustness," in *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 3571–3583, 2021. Available: <https://proceedings.neurips.cc/paper/2021/hash/1d49780520898fe37f0cd6b41c5311bf-Abstract.html>
- [6] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. CVPR*, pp. 113–123, 2019, doi: 10.1109/CVPR.2019.00020.
- [7] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 18613–18624, 2020. Available: <https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html>
- [8] D. Hendrycks et al., "AugMix: A simple data processing method to improve robustness and uncertainty," in *Proc. ICLR*, 2020. Available: <https://openreview.net/forum?id=S1gmrxHFvB>
- [9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018. Available: <https://openreview.net/forum?id=r1Ddp1-Rb>
- [10] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with Cutout," *arXiv:1708.04552*, 2017, doi: 10.48550/arXiv.1708.04552.
- [11] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 13001–13008, 2020, doi: 10.1609/aaai.v34i07.7000.
- [12] R. G. Lopes, D. Yin, B. Poole, J. Gilmer, and E. D. Cubuk, "Improving robustness without sacrificing accuracy with Patch Gaussian augmentation," *arXiv:1906.02611*, 2019, doi: 10.48550/arXiv.1906.02611.
- [13] Z. Huang et al., "IPMix: Label-preserving data augmentation method for training robust classifiers," in *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 63660–63673, 2023. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/c917d8b9e01427f3184d80ade22f4d1f-Abstract-Conference.html
- [14] N. Chen, J. Zhu, J. Chen, and B. Zhang, "Dropout training for support vector machines," in *Proc. AAAI Conf. Artif. Intell.*, vol. 28, no. 1, pp. 1752–1759, 2014, doi: 10.1609/aaai.v28i1.8965.
- [15] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Mach. Learn.*, vol. 52, no. 3, pp. 239–281, 2003, doi: 10.1023/A:1024068626366.
- [16] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979. Available: <https://www.jstor.org/stable/4615733>
- [17] E. Alpaydin and C. Kaynak, "Optical Recognition of Handwritten Digits," *UCI Mach. Learn. Repository*, 1998, doi: 10.24432/C50P49.
- [18] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011. Available: <https://jmlr.org/papers/v12/pedregosa11a.html>
- [19] P. Virtanen et al., "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nat. Methods*, vol. 17, pp. 261–272, 2020, doi: 10.1038/s41592-019-0686-2.
- [20] Scikit-learn developers, "load_digits," *scikit-learn documentation*. Available: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html (accessed Oct. 11, 2026).
- [21] D. Lakens, "Equivalence tests: A practical primer for t tests, correlations, and meta-analyses," *Soc. Psychol. Personal. Sci.*, vol. 8, no. 4, pp. 355–362, 2017, doi: 10.1177/1948550617697177.
- [22] Y. LeCun, C. Cortes, and C. J. C. Burges, "The MNIST database of handwritten digits." Available: <https://yann.lecun.org/exdb/mnist/> (accessed Oct. 11, 2026)

Mikhail D. Kashpirev is a student at the Institute of Information Technologies, MIREA - Russian Technological University, Moscow, Russian Federation. E-mail: imperiy05@mail.ru.