А. А. Трушкина
Московский технический университет связи и информатики (МТУСИ),
Москва, Россия

СЕЛЕКТИВНАЯ КЛАССИФИКАЦИЯ ТОВАРНЫХ ИЗОБРАЖЕНИЙ. ПИЛОТНАЯ ОЦЕНКА ПОКРЫТИЯ, РИСКА И ЭКОНОМИЧЕСКОГО ЭФФЕКТА

Аннотация

При внедрении моделей машинного обучения в системы управления мастер-данными (MDM) и товарной информацией (PIM) важно оценивать не только общую точность модели. Для бизнеса принципиально, какую долю объектов модель может обработать самостоятельно и сколько ошибок при этом останется незамеченными. В данной работе показатель Automation Rate интерпретируется как покрытие (coverage) селективного классификатора и рассматривается вместе с селективным риском (selective risk), обобщённым риском и калибровкой вероятностей.

Проведён ретроспективный анализ пилотного эксперимента с одним запуском пяти архитектур компьютерного зрения. ConvNeXt-Tiny показал наибольшее значение взвешенной F1-меры — 0,518 (95%-й bootstrap-интервал 0,486–0,549), тогда как ViT-Small/16 принял без ручной проверки наибольшую долю объектов при пороге уверенности 0,85: 26,9% против 14,2% у ConvNeXt-Tiny. Тем самым показано, что модель с наилучшей F1-мерой не обязательно обеспечивает наибольшее покрытие. Однако эти результаты нельзя считать доказательством преимущества конкретной архитектуры, поскольку модели не калибровались и не обучались повторно с разными случайными инициализациями.

В качестве методического результата предложен формализованный протокол подтверждающего эксперимента. Он предусматривает разбиение данных без пересечения товарных идентичностей, отдельные выборки для калибровки, выбора порога, независимой проверки риска и финального тестирования, пять запусков каждой архитектуры, температурное масштабирование и расчёт ECE, NLL, Brier score, AURC и AUGRC. Размер сертификационной выборки предлагается определять заранее по ожидаемому покрытию, допустимому риску и требуемой вероятности успешной проверки. Выборка из 400 объектов рассматривается только как предварительная оценка процедуры. Экономическая модель учитывает ручную проверку, стоимость ошибочных автоматических решений, текущие расходы и первоначальные инвестиции. В иллюстративном сценарии предпочтение между ViT-Small/16 и ConvNeXt-Tiny меняется при цене ошибки около 4 262 руб. Все численные выводы относятся к пилотным данным; подтверждающие выводы потребуют полного многократного эксперимента и сертификационной выборки достаточного объёма.

Ключевые слова: селективная классификация; покрытие (coverage); селективный риск; калибровка вероятностей; сертификация риска; AURC; AUGRC; контур с участием оператора; PIM; MDM; компьютерное зрение; экономическая эффективность ИИ.

1. Введение

Системы управления мастер-данными (Master Data Management, MDM) и товарной информацией (Product Information Management, PIM) отвечают за согласованность справочников, наполнение товарных карточек и передачу данных в каналы продаж. В крупных каталогах категоризация товаров является масштабной и часто мультимодальной задачей: на результат влияют изображения, тексты, атрибуты, иерархия категорий, шум разметки и изменения таксономии [14–17]. При росте потока ручная классификация становится узким местом. При этом обычные Accuracy и F1 не отвечают на практический вопрос: какую часть карточек можно передать модели без проверки оператором.

На практике такие системы обычно работают по схеме «модель + оператор». Модель автоматически обрабатывает только те объекты, в которых достаточно уверена, а остальные передаёт человеку. Поэтому важны два показателя: доля автоматически принятых решений — покрытие (coverage) — и доля ошибок среди этих решений — селективный риск (selective risk) [8–10]. Термин

Automation Rate в данной работе используется лишь как возможное бизнес-наименование покрытия, а не как новая метрика машинного обучения.

Есть и вторая проблема: одинаковое значение softmax-вероятности у разных моделей не означает одинаковую надёжность. Нейронные сети могут быть переуверенными, поэтому один и тот же порог уверенности способен давать разные фактические риски [11]. По этой причине сравнивать модели только по покрытию при фиксированном пороге некорректно. Необходимо дополнительно оценивать калибровку вероятностей или выбирать рабочий порог на отдельной выборке после калибровки.

Третья проблема связана с экономикой процесса. Рост покрытия уменьшает объём ручной работы, но может увеличить число ошибочных решений, которые попадут в систему без проверки. Цена такой ошибки зависит от конкретного процесса и может включать повторную обработку, исправление опубликованной карточки, санкции площадки и потери выручки. Поэтому модель следует выбирать не по одному техническому показателю, а с учётом стоимости ручной проверки и стоимости ошибочного автоматического решения.

Цель работы — предложить формализованную методику, которая связывает показатели селективной классификации с операционной и экономической эффективностью PIM-процесса, а также корректно оценить имеющиеся пилотные результаты.

Для достижения цели поставлены следующие задачи:

– сопоставить бизнес-показатель автоматической обработки со стандартным понятием покрытия в селективной классификации;

– определить покрытие, селективный и обобщённый риск, AURC и AUGRC;

– включить в методику калибровку вероятностей и независимую проверку выбранного порога;

– сформировать протокол с независимым тестированием и повторными запусками моделей;

– сравнить вспомогательный композитный индекс с экономическим критерием выбора модели;

– выполнить ретроспективный анализ пилотных результатов, не приписывая исследованию отсутствующие измерения.

Научная новизна работы состоит не во введении новой метрики, а в прикладной адаптации методов селективной классификации к процессам PIM/MDM. Предлагаемый подход разделяет калибровку модели, выбор рабочей точки и независимую проверку риска, заранее планирует необходимый объём сертификационной выборки и связывает технические показатели с прозрачной экономической моделью. Методический результат при этом отделён от ограниченных пилотных данных.

2. Связанные исследования

2.1. Категоризация товаров и архитектуры компьютерного зрения

DeepFashion2 объединяет коммерческие и пользовательские изображения одежды, допускает несколько предметов на одном снимке и связывает изображения одного товара через идентификаторы pair_id и style [1]. Поэтому единицей классификации должен быть отдельный аннотированный объект, а связанные изображения одного товара не должны попадать в разные выборки. В пилотном эксперименте сравнивались ResNet50 [2], EfficientNet-B3 [3], ViT-Small/16 [4], Swin-Tiny [5] и ConvNeXt-Tiny [6]. Набор включает классические свёрточные сети, визуальные трансформеры и современные CNN. Вместе с тем DeepFashion2 остаётся исследовательским набором данных, а не точной копией промышленного PIM-потока, поэтому перенос результатов на реальные каталоги требует отдельной проверки.

Grad-CAM [7] помогает увидеть, на какие области изображения опиралась модель, но сам по себе не доказывает преимущество архитектуры и не заменяет количественную оценку неопределённости. Исследования в электронной коммерции также показывают пользу текстовых и

структурных признаков, мультимодального предобучения и явного учёта товарной таксономии [14–17]. Поэтому ошибки чисто визуальной модели следует рассматривать как основание для будущего мультимодального эксперимента, а не как принципиальный предел качества.

2.2. Селективная классификация и отказ от автоматического решения

Идея классификации с возможностью отказа от решения появилась задолго до современных нейронных сетей [8]. Селективный классификатор состоит из обычной модели и правила отбора: система принимает предсказание только при достаточной уверенности, а в остальных случаях передаёт объект на ручную проверку. Geifman и El-Yaniv показали, что такой контур следует оценивать по компромиссу между риском и покрытием [9]. В SelectiveNet отказ от решения обучается вместе с классификатором [10], однако в прикладных пилотах часто используют более простую схему — порог по уверенности уже обученной модели.

Одну рабочую точку описывают покрытие и селективный риск. Чтобы сравнить модели по всему диапазону порогов, строят кривую «риск–покрытие» и рассчитывают площадь под ней — AURC. Более поздняя работа указала на ограничения AURC и предложила AUGRC — площадь под кривой обобщённого риска, которая непосредственно отражает средний риск необнаруженной ошибки [12]. В данном исследовании используются обе метрики: AURC сохраняет сопоставимость с предыдущими работами, а AUGRC дополняет её более понятной интерпретацией скрытых ошибок.

2.3. Калибровка вероятностей

Калиброванная модель должна выдавать такие вероятности, которые соответствуют наблюдаемой частоте правильных ответов. Например, среди объектов с уверенностью около 0,8 примерно 80% предсказаний должны быть верными. Guo и соавторы показали, что современные глубокие сети часто переоценивают свою уверенность, а температурное масштабирование (temperature scaling) позволяет исправить шкалу вероятностей без переобучения модели [11]. Поэтому для сравнения архитектур недостаточно одного покрытия: необходимо сообщать ECE, NLL и Brier score до и после калибровки.

Температурное масштабирование не меняет выбранный класс и, следовательно, не влияет на Accuracy и F1. Оно меняет только численные значения уверенности. Поэтому покрытие при фиксированном пороге может измениться, даже если все предсказания остались прежними. Рабочий порог следует выбирать после калибровки и на данных, которые не использовались при обучении модели.

2.4. Экономический эффект контура с участием оператора

Экономический эффект возникает не из самого значения F1, а из того, как модель меняет рабочий процесс. В PIM/MDM стоимость ручной проверки и стоимость ошибочной категоризации могут различаться на порядки. Поэтому модель с более высоким покрытием не обязательно оказывается дешевле. В работе используется ожидаемая стоимость, которая включает ручную проверку, ошибочные автоматические решения, текущие расходы на ИИ-модуль и первоначальные инвестиции. Все параметры задаются явно и анализируются как сценарные, а не как универсальные отраслевые нормы.

3. Методика исследования

3.1. Классификатор и оценка уверенности

Пусть $f_\theta: \mathcal{X} \to \mathbb{R}^C$ — модель, выдающая вектор логитов для C классов. Вероятности после temperature scaling определяются как

$$p_T(y = c \mid x) = \frac{\exp\left(f_{\theta,c}(x)/T\right)}{\sum_{j=1}^{C} \exp\left(f_{\theta,j}(x)/T\right)}, \qquad T > 0.$$

Предсказанный класс и функция уверенности имеют вид

$$\hat{y}(x) = \arg\max_c p_T(y = c \mid x), \qquad s(x) = \max_c p_T(y = c \mid x).$$

При $T = 1$ используются исходные softmax-вероятности. Параметр T оценивается на calibration-выборке минимизацией negative log-likelihood.

3.2. Покрытие и селективный риск

Для порога τ функция отбора определяется как

$$g_\tau(x) = \mathbb{I}[s(x) \geq \tau].$$

Если уверенность модели не ниже заданного порога, решение принимается автоматически; в противном случае объект направляется на ручную проверку. Эмпирическое покрытие на выборке из n объектов определяется как

$$\widehat{\mathrm{Cov}}(\tau) = \frac{1}{N}\sum_{i=1}^{N} g_\tau\,(x_i).$$

Доля объектов, принятых автоматически, соответствует показателю coverage — «покрытие». Обозначение Automation Rate может использоваться как понятное бизнес-наименование, но в математическом описании применяется стандартный термин.

При 0/1-функции ошибки $\ell_i = \mathbb{I}[\hat{y}_i \neq y_i]$ selective risk определяется как

$$\hat{R}_{\mathrm{sel}}(\tau) = \frac{\sum_{i=1}^{N} \ell_i\, g_\tau(x_i)}{\sum_{i=1}^{N} g_\tau\,(x_i)}.$$

Если модель не принимает автоматически ни одного решения, селективный риск математически не определён. Такая рабочая точка исключается из анализа из-за нулевого покрытия.

Обобщённый риск показывает долю ошибочных автоматических решений во всём входном потоке:

$$\hat{R}_{\mathrm{gen}}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \ell_i\, g_\tau(x_i) = \widehat{\mathrm{Cov}}(\tau)\hat{R}_{\mathrm{sel}}(\tau).$$

3.3. Кривая «риск–покрытие» и показатели AURC/AUGRC

При изменении порога меняются и покрытие, и риск. Их совместную динамику отражает кривая «риск–покрытие». Для сравнения моделей по всему диапазону порогов используются следующие показатели:

$$\mathrm{AURC} = \int_0^1 R_{\mathrm{sel}}\,(u)\,du,$$

$$\mathrm{AUGRC} = \int_0^1 R_{\mathrm{gen}}\,(u)\,du,$$

где c — покрытие. Чем меньше AURC и AUGRC, тем лучше уверенность модели отделяет правильные ответы от ошибочных. AURC приводится для сопоставимости с предыдущими исследованиями, а AUGRC — как показатель среднего риска необнаруженной ошибки [12].

3.4. Оценка калибровки

Отрицательное логарифмическое правдоподобие (NLL) рассчитывается как

$$\mathrm{NLL} = -\frac{1}{N}\sum_{i=1}^{N} \log p_T(y_i \mid x_i).$$

Для многоклассовой задачи показатель Brier score имеет вид

$$\mathrm{Brier} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} (p_T(y = c \mid x_i) - \mathbb{I}[y_i = c])^2.$$

Expected calibration error при разбиении confidence на M интервалов вычисляется как

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|.$$

ECE зависит от того, как диапазон уверенности разбит на интервалы. Поэтому вместе с ним сохраняются NLL, Brier score и данные для диаграмм надёжности, которые позволяют визуально оценить соответствие уверенности фактической точности.

3.5. Выбор и независимая проверка рабочей точки

Порог 0,85 не рассматривается как универсальное значение. После калибровки на модельно-валидационной выборке выбирается кандидатный порог, который обеспечивает максимальное покрытие при заданном ограничении на селективный риск:

$$\tau^* = arg\ max\tau\ \hat{C}ovval(\tau),\ \ npu\ \ \hat{R}sel,val(\tau) \leq rmax.$$
$$U1-\delta(kcert,\ ncert) \leq rmax.$$

После выбора порог фиксируется и только один раз проверяется на независимой сертификационной выборке. Для числа ошибочных автоматически принятых решений kcert и общего числа принятых решений ncert рассчитывается односторонняя точная верхняя граница Клоппера–Пирсона. Рабочая точка считается подтверждённой, только если эта граница не превышает допустимый риск rmax. В исследовании используется rmax = 0,05 при доверительном уровне 0,95. Если условие не выполнено, система не принимает решения автоматически. Повторный подбор порога на тех же сертификационных данных запрещён. Окончательная пара «архитектура — обученная модель — порог» должна быть выбрана до обращения к сертификационной выборке. Если одновременно проверяются несколько кандидатов, необходима поправка на множественные проверки либо независимые сертификационные части. Такой порядок соответствует логике Learn-Then-Test [13].

3.6. Планирование объёма сертификационной выборки

Размер сертификационной выборки определяется прежде всего числом объектов, которые модель примет автоматически. Приближённо ncert ≈ Ncert·coverage, где Ncert — общий объём выборки. При допустимом риске 5% и одностороннем доверительном уровне 95% верхняя граница Клоппера–Пирсона не превышает 5%, например, при следующих сочетаниях числа принятых решений и ошибок: 59 и 0; 93 и 1; 124 и 2; 153 и 3; 181 и 4.

Из этого следует, что выборка из 400 объектов слишком мала для строгой проверки при наблюдаемом пилотном покрытии. При покрытии 14,2% модель примет около 57 объектов, и риск не удастся подтвердить даже при отсутствии ошибок. При покрытии 26,9% будет принято около 108 объектов, поэтому проверку можно пройти лишь при нуле или одной ошибке. Таким образом, 400 объектов подходят только для технической проверки контура, но не для окончательного статистического вывода.

Необходимый объём следует рассчитывать заранее с учётом предполагаемого истинного риска и требуемой вероятности успешной сертификации. Например, если истинный селективный риск равен 2%, для вероятности успешной проверки не менее 80% нужно как минимум 234 принятых решения, а для вероятности 90% — 311. При покрытии 14,2% это соответствует общей выборке примерно из 1 648 и 2 191 объектов; при покрытии 26,9% — из 870 и 1 157 объектов. Объём фиксируется до просмотра результатов моделей. Если benchmark не позволяет сформировать такую выборку, проверку следует проводить на отдельном независимом потоке производственных данных; иначе результат необходимо обозначить как предварительный.

3.7. Вспомогательный композитный показатель

Для анализа чувствительности ранжирования рассчитывается композитный показатель

$$J_\alpha(m) = \alpha\widetilde{F1}(m) + (1-\alpha)\widetilde{\mathrm{Cov}}(m),$$

Тильда обозначает min–max-нормализацию по набору сравниваемых моделей. Показатель используется только для анализа чувствительности: его значение зависит от состава моделей и

выбранного порога и не учитывает цену ошибки. Для практического выбора основной считается экономическая функция затрат.

3.8. Экономическая модель

Для месячного потока N ожидаемые затраты гибридного процесса задаются как

$$C_{\mathrm{AI}} = C_{\mathrm{review}} N_{\mathrm{review}} + C_{\mathrm{error}} N_{\mathrm{wrong\,auto}} + OPEX_{\mathrm{AI}},$$

$$N_{\mathrm{review}} = N(1 - \mathrm{Cov}), \qquad N_{\mathrm{wrong\,auto}} = N\,\mathrm{Cov}\,R_{\mathrm{sel}}.$$

Базовые затраты полностью ручного процесса равны

$$C_{\mathrm{manual}} = N\,t_{\mathrm{manual}} c_h,$$

где время выражается в часах, а c_h — стоимость человеко-часа. Ежемесячный эффект определяется как $B = C_{\mathrm{manual}} - C_{\mathrm{AI}}$. При наличии начальных инвестиций $CAPEX$:

$$\mathrm{Payback} = \frac{CAPEX}{B}, \qquad \mathrm{ROI}_{1y} = \frac{12B - CAPEX}{CAPEX}.$$

Если $B \le 0$, срок окупаемости не определён. Такой расчёт исключает двойное вычитание OPEX: стоимость модуля учитывается только в C_{AI}.

4. План эксперимента

4.1. Доступные данные и ограничения пилота

Доступные материалы содержат итоговые метрики одного запуска каждой архитектуры, матрицы ошибок и графики. Индивидуальные значения уверенности и исходные логиты не сохранились, поэтому невозможно восстановить полный набор калибровочных и селективных метрик либо оценить изменчивость результатов между повторными запусками.

В работе используются только результаты, которые можно проверить по имеющимся материалам. Раздел 5 представляет ретроспективный анализ пилота. Полный подтверждающий эксперимент рассматривается как следующий этап исследования, поэтому статья позиционируется как методическое пилотное исследование, а не как завершённое сравнение архитектур.

4.2. Данные и формирование выборок

Пять классов были выбраны как наиболее представленные в размеченной части DeepFashion2. В пилотный анализ включены 5 000 объектов, из которых 1 000 составили тестовую выборку. Сохранившиеся материалы не позволяют однозначно установить, использовались ли во всех случаях вырезки отдельных предметов или полные изображения; поэтому результаты интерпретируются как предварительные. В подтверждающем эксперименте единицей классификации должен быть один аннотированный предмет, выделенный по bounding box.

Для подтверждающего исследования 4 000 объектов из размеченной обучающей части при фиксированном seed разбиения 2026 делятся методом StratifiedGroupKFold на десять непересекающихся по товарной идентичности частей: семь используются для обучения, по одной — для калибровки, модельной валидации и предварительной сертификации. Ещё 1 000 объектов из размеченной официальной validation-части резервируются для финального теста. Тестовые объекты не должны пересекаться с пилотной выборкой и не должны использоваться при разработке методики или выборе конфигурации. Если это условие нельзя выполнить на DeepFashion2, необходима новая внешняя выборка либо независимые данные из реального PIM-потока. Неразмеченная официальная test-часть для локального расчёта метрик не используется.

Таблица 1 — Функциональное группово-непересекающееся разделение выборок

Выборка	Ориентировочный объём	Назначение
Обучающая	≈ 2 800	обучение параметров модели
Калибровочная	≈ 400	настройка температурного масштабирования
Модельно-валидационная	≈ 400	ранняя остановка и выбор кандидатного порога

Выборка	Ориентировочный объём	Назначение
Сертификационная	≈ 400 (пилот)	однократная независимая проверка риска
Тестовая	1 000	единственная финальная оценка

Сертификационная часть объёмом около 400 объектов предназначена только для предварительной оценки процедуры. В полном исследовании её заменяют выборкой, размер которой рассчитан заранее по процедуре раздела 3.6 и не меняется после просмотра результатов моделей.

Основной идентификатор группы — pair_id или другой доступный идентификатор товара, дополненный категорией. Такое разбиение не позволяет изображениям одной commercial-consumer пары и повторным записям одного объекта попасть в разные внутренние выборки. Дополнительно исключаются дублирующиеся изображения. Если товарный идентификатор отсутствует, допускается резервное группирование по изображению, но в этом случае отсутствие утечки по идентичности не утверждается. При наличии bounding box модель получает вырезанный фрагмент с одним аннотированным предметом. Повреждённые изображения исключаются из выборки.

Рисунок 1 — Схема подтверждающего эксперимента: выбор порога и независимая проверка риска

4.3. Сравниваемые архитектуры и обучение

Таблица 2 — Архитектуры, включённые в сравнительный эксперимент

Архитектура	Тип модели	Параметров, млн	Роль в сравнении
ResNet50	классическая CNN	25,6	базовая свёрточная модель
EfficientNet-B3	масштабируемая CNN	12,2	компактная модель
ViT-Small/16	визуальный трансформер	22,1	глобальное внимание
Swin-Tiny	иерархический трансформер	28,3	оконное внимание
ConvNeXt-Tiny	современная CNN	28,6	современная свёрточная модель

Все архитектуры инициализируются весами ImageNet и обучаются по одному заранее зафиксированному протоколу: не более 40 эпох, оптимизатор AdamW, начальная скорость обучения $5 \cdot 10^{-5}$, weight decay 10^{-4}, batch size 16 и label smoothing 0,1. В первые 10 эпох обучается только классификационная голова, затем размораживается основная часть сети. Ранняя остановка выполняется по loss на модельно-валидационной выборке. Используются AdamW [18] и расписание OneCycle [19]. RandAugment [20] может применяться только как заранее заявленный дополнительный вариант и не смешивается с основным сравнением.

Для всех архитектур используется одинаковая базовая конфигурация. Дополнительный поиск гиперпараметров, если он проводится, завершается до обращения к тестовым данным и описывается отдельно. Перед оценкой восстанавливаются наилучшие веса конкретной пары «архитектура — случайная инициализация». В подтверждающее сравнение также включаются две простые контрольные модели: классификатор большинства и линейный классификатор на замороженных

ImageNet-признаках. Это позволяет понять, насколько сложные нейронные сети действительно превосходят простые базовые решения.

4.4. Повторные запуски и оценка неопределённости

Каждая архитектура обучается пять раз с seed 17, 29, 42, 73 и 101. Состав выборок при этом остаётся неизменным, поэтому различия между запусками отражают изменчивость обучения, а не случайное перераспределение данных. По пяти запускам рассчитываются среднее, стандартное отклонение и 95%-й t-интервал. Bootstrap на уровне тестовых объектов может дополнительно оценить неопределённость конечной выборки, но не заменяет повторное обучение моделей.

Для пилотных матриц ошибок доступны только агрегированные частоты, поэтому использован параметрический multinomial bootstrap с 20 000 повторов. Полученные интервалы описывают неопределённость тестовой выборки при фиксированной модели, но не учитывают изменчивость обучения. Перекрытие интервалов не интерпретируется как формальное доказательство эквивалентности моделей.

5. Результаты ретроспективного анализа

5.1. Качество классификации и покрытие

В таблице 3 приведены точечные оценки и 95%-е bootstrap-интервалы, восстановленные по матрицам ошибок. Пилотная тестовая выборка включала 300 объектов класса short sleeve top, 244 — trousers, 113 — shorts, 156 — long sleeve top и 187 — skirt. Интервалы для покрытия не рассчитывались, поскольку исходные логиты и индивидуальные значения уверенности не сохранились. Показатель mAP не включён: без исходных вероятностей невозможно проверить его определение и расчёт. Значения задержки также не публикуются, поскольку исходная процедура измерения не была достаточно строго зафиксирована.

Таблица 3 — Пилотные показатели качества и некалиброванного покрытия

Архитектура	Accuracy [95% ДИ]	Взвешенная F1 [95% ДИ]	Покрытие при τ=0,85
ResNet50	0,452 [0,421; 0,483]	0,445 [0,414; 0,477]	5,0%
EfficientNet-B3	0,384 [0,354; 0,415]	0,379 [0,349; 0,410]	7,3%
ViT-Small/16	0,486 [0,456; 0,517]	0,480 [0,449; 0,512]	26,9%
Swin-Tiny	0,459 [0,428; 0,491]	0,456 [0,425; 0,488]	15,5%
ConvNeXt-Tiny	0,522 [0,491; 0,553]	0,518 [0,486; 0,549]	14,2%

Наибольшую взвешенную F1-меру показала ConvNeXt-Tiny. Однако её доверительный интервал заметно перекрывается с интервалом ViT-Small/16. Это не доказывает равенство моделей, но показывает, что разницу одного запуска нельзя считать устойчивым преимуществом без повторного обучения.

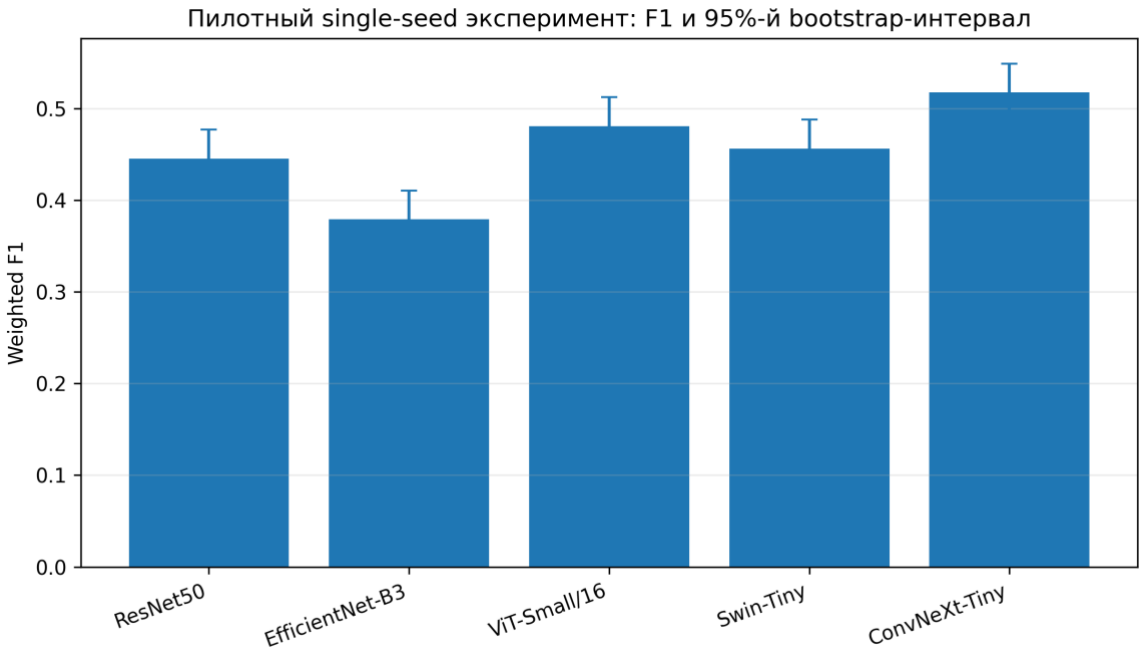

Рисунок 2 — Взвешенная F1-мера и 95%-е bootstrap-интервалы пилотных оценок

При пороге 0,85 наибольшее покрытие показала ViT-Small/16. Однако этот результат одновременно зависит от способности модели ранжировать правильные и ошибочные ответы и от масштаба её softmax-вероятностей. Без калибровочных метрик и температурного масштабирования нельзя определить, связано ли высокое покрытие с действительно полезной уверенностью или с переуверенностью модели.

Рисунок 3 — Доля объектов, принятых автоматически при пороге 0,85

5.2. Чувствительность композитного показателя

После min–max-нормализации пилотных результатов значение композитного показателя зависит от веса α. При α = 0,5 ViT-Small/16 получает 0,866, а ConvNeXt-Tiny — 0,710. При α = 0,7 значения составляют 0,813 и 0,826 соответственно.

Таблица 4 — Чувствительность вспомогательного ранжирования к весу F1

Архитектура	F1, норм.	Покр., норм.	J(α=0,3)	J(α=0,5)	J(α=0,7)
ResNet50	0,479	0,000	0,144	0,239	0,335
EfficientNet-B3	0,000	0,105	0,074	0,053	0,032
ViT-Small/16	0,733	1,000	0,920	0,866	0,813
Swin-Tiny	0,556	0,479	0,502	0,518	0,533
ConvNeXt-Tiny	1,000	0,420	0,594	0,710	0,826

При увеличении веса F1 лидером становится ConvNeXt-Tiny. Кроме того, min–max-нормализация изменится, если в сравнение добавить новую слабую или сильную модель. Поэтому показатель J используется только для анализа чувствительности и не рассматривается как самостоятельный критерий промышленного или инвестиционного решения.

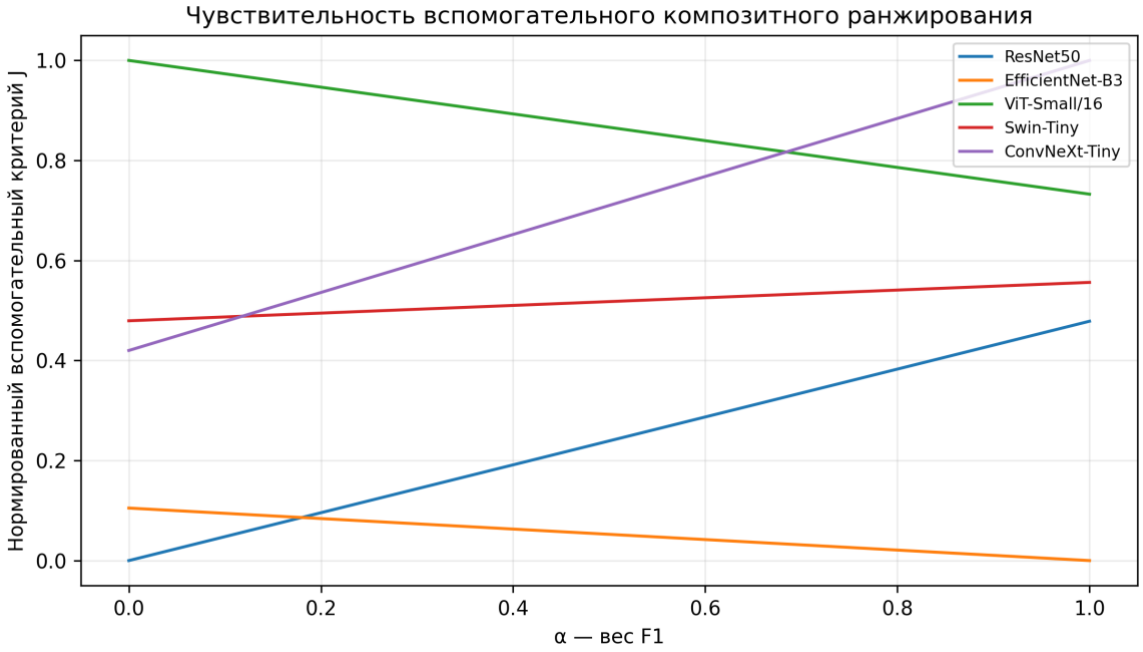

Рисунок 4 — Влияние веса F1 на композитное ранжирование моделей

5.3. Анализ ошибок

Матрица ошибок ConvNeXt-Tiny показывает, что модель часто путает визуально близкие классы: short sleeve top с long sleeve top и shorts с trousers. Причиной могут быть кадрирование и отсутствие текстовых атрибутов, но пилотные данные не позволяют связать ошибки с конкретной архитектурой или считать их принципиально неустранимыми. Для проверки нужны объектные фрагменты, разбиение по товарной идентичности и сравнение с мультимодальными моделями.

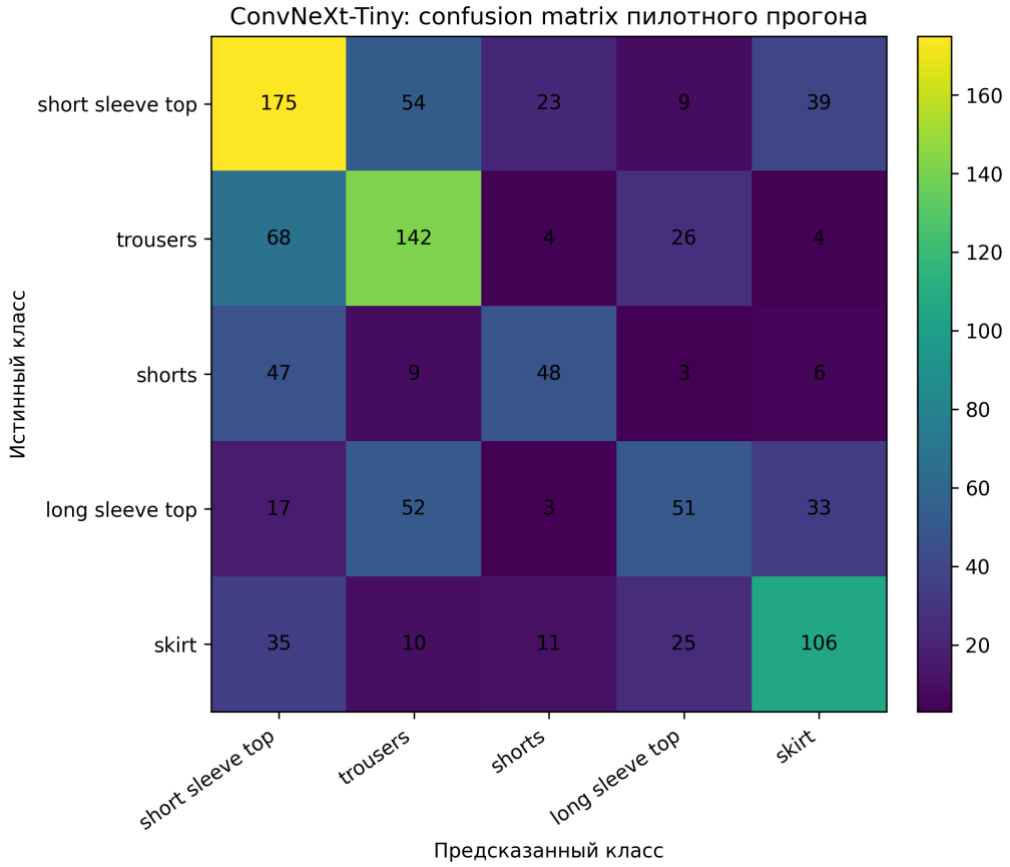

Рисунок 5 — Матрица ошибок ConvNeXt-Tiny в пилотном эксперименте

5.4. Иллюстративная экономическая оценка

В пилотных материалах селективный риск при пороге 0,85 сохранён только для ViT-Small/16 (3,8%) и ConvNeXt-Tiny (5,1%). Проверить происхождение этих значений без исходных логитов невозможно, поэтому они используются исключительно как параметры иллюстративного сценария. Предполагается поток 3 000 карточек в месяц, 10 минут на полностью ручную обработку, 5 минут на проверку отклонённого моделью объекта, стоимость труда 1 200 руб./ч, текущие расходы ИИ-модуля 150 тыс. руб./мес., первоначальные инвестиции 900 тыс. руб. и стоимость одной ошибочной автоматической классификации 3 000 руб. Эти значения не претендуют на роль рыночных нормативов.

Таблица 5 — Иллюстративные операционные и стоимостные показатели при цене ошибки 3 000 руб.

Модель	Покрытие, %	Риск, %	Ручная проверка, шт./мес.	Ошибки, шт./мес.	Затраты, руб./мес.	Эффект, руб./мес.
ViT-S/16	26,9	3,8	2 193	30,7	461 298	138 702
ConvNeXt	14,2	5,1	2 574	21,7	472 578	127 422

В таблице покрытие и риск приведены в процентах, объём ручной проверки и число ошибок — в объектах за месяц, а затраты и эффект — в рублях за месяц. Расчётный ROI первого года составляет 84,9% для ViT-Small/16 и 69,9% для ConvNeXt-Tiny; срок окупаемости — 6,5 и 7,1 месяца соответственно. При заданных предположениях ViT-Small/16 требует меньших ежемесячных затрат, поскольку сокращает объём ручной проверки. Одновременно она создаёт больше ошибочных автоматических решений: около 30,7 против 21,7 в месяц. Расчётная точка безразличия между моделями составляет примерно 4 262 руб. за ошибку. Это значение зависит от пилотных покрытия и риска и не переносится автоматически на другие процессы. После полного эксперимента экономический расчёт следует повторить для каждого запуска, а неопределённость стоимости оценить по изменчивости рабочих точек.

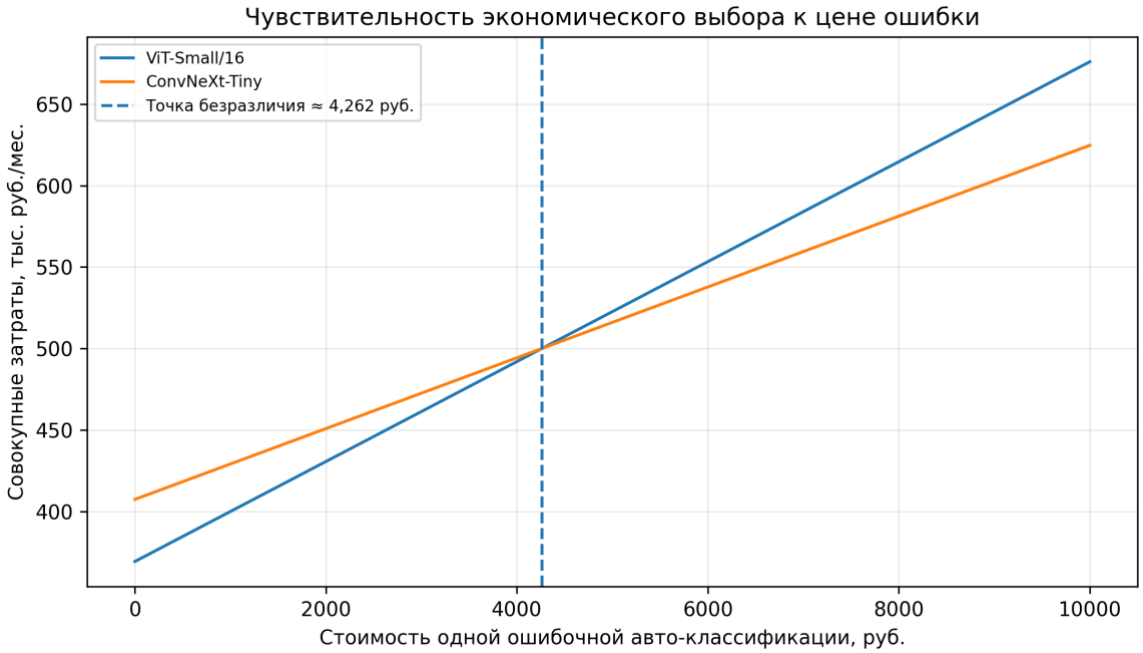

Рисунок 6 — Влияние стоимости ошибки на экономический выбор модели

Таким образом, пилот не подтверждает универсальное экономическое преимущество ViT-Small/16. Выбор зависит от сочетания покрытия, селективного риска и параметров конкретного бизнес-процесса. Экономическая модель дополняет, но не заменяет независимую оценку качества.

6. Обсуждение результатов

Пилотный эксперимент показывает важное расхождение между общим качеством классификации и долей решений, которые модель считает достаточно уверенными. Более высокое покрытие ViT-Small/16 при пороге 0,85 не означает, что механизм self-attention формирует более надёжную уверенность. Для такого вывода нужны повторные запуски, калибровка, полные кривые «риск–покрытие» и независимая проверка рабочей точки. Пока архитектурное объяснение остаётся гипотезой.

Покрытие полезно как операционный показатель, но его нельзя максимизировать отдельно от риска. При одинаковом покрытии модели могут совершать разное число ошибочных автоматических решений; при одинаковом селективном риске — передавать оператору разный объём работы. Обобщённый риск показывает долю скрытых ошибок во всём потоке, а экономическая функция переводит технические различия в затраты конкретного процесса.

Калибровка вероятностей и качество селективного ранжирования — разные свойства. Температурное масштабирование может улучшить ECE и NLL, но как монотонное преобразование логитов не обязано улучшать AURC и AUGRC. Поэтому калибровочные и селективные метрики следует представлять раздельно: первые описывают достоверность численной вероятности, вторые — способность уверенности отделять правильные ответы от ошибочных.

Экономическая модель также показывает, почему нельзя назвать одну архитектуру лучшей вне контекста. При низкой цене ошибки важнее сократить ручную проверку, а при высокой — уменьшить число необнаруженных ошибок. Для реального внедрения параметры затрат должны оцениваться по данным конкретного процесса и журналу инцидентов. Анализ чувствительности должен учитывать объём потока, стоимость труда, цену ошибки, OPEX, CAPEX и возможный дрейф данных.

У пилотной части есть существенные ограничения: один benchmark, пять классов, небольшой объём данных, один запуск и отсутствие исходных логитов. Bootstrap-интервалы не отражают изменчивость обучения. DeepFashion2 содержит несколько предметов на изображении и связанные commercial-consumer пары, поэтому обычное построчное разбиение может приводить к утечке; это ограничение учитывается в плане подтверждающего эксперимента. Выборка из 400 объектов

недостаточна для строгой сертификации риска 5% при наблюдаемом покрытии и используется только для предварительной оценки процедуры. Для окончательного эксперимента нужен заранее рассчитанный объём или независимый производственный поток. Финансовые параметры также являются сценарными. Наконец, чисто визуальная классификация упрощает реальную PIM-задачу, где обычно доступны название, бренд, атрибуты и иерархия категорий.

7. Заключение

Показатель Automation Rate в рамках селективной классификации интерпретируется как покрытие. Само по себе покрытие недостаточно: его необходимо анализировать вместе с селективным и обобщённым риском и с качеством калибровки. В работе описаны AURC, AUGRC, экономическая модель контура с участием оператора и процедура, которая отделяет выбор порога от независимой проверки риска.

Получены три основных результата. Во-первых, модель с наибольшей F1-мерой не совпала с моделью, обеспечившей наибольшее покрытие: ConvNeXt-Tiny лидировала по взвешенной F1-мере, тогда как ViT-Small/16 — по некалиброванному покрытию. Во-вторых, сравнение покрытия при одном фиксированном пороге без калибровки не позволяет сопоставлять надёжность разных архитектур. В-третьих, экономически предпочтительная модель зависит от стоимости ошибки: расчётная точка безразличия около 4 262 руб. относится только к заданному сценарию.

Предложен формализованный протокол подтверждающего эксперимента с группово-непересекающимся разбиением, пятью запусками каждой архитектуры, температурным масштабированием и отдельными выборками для выбора и сертификации порога. Показано, что размер сертификационной выборки необходимо планировать заранее; 400 объектов достаточно лишь для предварительной оценки процедуры. В текущем виде работа представляет завершённое пилотное методическое исследование. Для строгих сравнительных выводов требуется полный многократный эксперимент и статистически достаточная независимая сертификационная выборка.

Список литературы

1. Ge Y., Zhang R., Wang X., Wu L., Tang X., Luo P. DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images // Proceedings of CVPR. 2019. P. 5337–5345.

2. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition // Proceedings of CVPR. 2016. P. 770–778. DOI: 10.1109/CVPR.2016.90.

3. Tan M., Le Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks // Proceedings of ICML. 2019. P. 6105–6114.

4. Dosovitskiy A., Beyer L., Kolesnikov A. et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale // ICLR. 2021.

5. Liu Z., Lin Y., Cao Y. et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows // Proceedings of ICCV. 2021. P. 10012–10022.

6. Liu Z., Mao H., Wu C.-Y. et al. A ConvNet for the 2020s // Proceedings of CVPR. 2022. P. 11976–11986.

7. Selvaraju R. R., Cogswell M., Das A. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization // Proceedings of ICCV. 2017. P. 618–626. DOI: 10.1109/ICCV.2017.74.

8. Chow C. K. An Optimum Character Recognition System Using Decision Functions // IRE Transactions on Electronic Computers. 1957. Vol. EC-6, No. 4. P. 247–254.

9. Geifman Y., El-Yaniv R. Selective Classification for Deep Neural Networks // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 4878–4887.

10. Geifman Y., El-Yaniv R. SelectiveNet: A Deep Neural Network with an Integrated Reject Option // Proceedings of ICML. PMLR. 2019. Vol. 97. P. 2151–2159.

11. Guo C., Pleiss G., Sun Y., Weinberger K. Q. On Calibration of Modern Neural Networks // Proceedings of ICML. PMLR. 2017. Vol. 70. P. 1321–1330.

12. Traub J., Bungert T. J., Lüth C. T. et al. Overcoming Common Flaws in the Evaluation of Selective Classification Systems // Advances in Neural Information Processing Systems. 2024. Vol. 37. P. 2323–2347.

13. Angelopoulos A. N., Bates S., Candès E. J., Jordan M. I., Lei L. Learn Then Test: Calibrating Predictive Algorithms to Achieve Risk Control // The Annals of Applied Statistics. 2025. Vol. 19, No. 2. DOI: 10.1214/24-AOAS1998.

14. Dong X., Zhan X., Wu Y., Wei Y., Kampffmeyer M. C., Wei X., Lu M., Wang Y., Liang X. M5Product: Self-Harmonized Contrastive Learning for E-Commercial Multi-Modal Pretraining // Proceedings of CVPR. 2022. P. 21252–21262.

15. Jin Y., Li Y., Yuan Z., Mu Y. Learning Instance-Level Representation for Large-Scale Multi-Modal Pretraining in E-Commerce // Proceedings of CVPR. 2023. P. 11060–11069.

16. Gong S., Zhou Z., Wang S. et al. Transferable and Efficient: Unifying Dynamic Multi-Domain Product Categorization // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track). 2023. P. 476–486. DOI: 10.18653/v1/2023.acl-industry.46.

17. Chen S., Bouadjenek M. R., Naseem U. et al. Leveraging Taxonomy and LLMs for Improved Multimodal Hierarchical Classification // Proceedings of the 31st International Conference on Computational Linguistics. 2025. P. 6244–6254.

18. Loshchilov I., Hutter F. Decoupled Weight Decay Regularization // ICLR. 2019.

19. Smith L. N., Topin N. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates // Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. 2019. Vol. 11006.

20. Cubuk E. D., Zoph B., Shlens J., Le Q. V. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space // Proceedings of CVPR Workshops. 2020.

Information about the article in English

Title: Selective Classification of Product Images: A Pilot Evaluation of Coverage, Risk, and Economic Impact.

Author: A. A. Trushkina, Moscow Technical University of Communications and Informatics, Moscow, Russia.

Abstract. When machine-learning models are deployed in Master Data Management and Product Information Management systems, overall classification accuracy is not sufficient. A practical system must also estimate what share of incoming items can be processed automatically and how many errors may remain undetected. This study interprets Automation Rate as selective-classification coverage and evaluates it together with selective risk, generalized risk, probability calibration, and operational cost. A retrospective single-seed pilot across five vision architectures found the highest weighted F1 for ConvNeXt-Tiny (0.518; bootstrap 95% interval 0.486–0.549) and the highest uncalibrated coverage at a confidence threshold of 0.85 for ViT-Small/16 (26.9% versus 14.2%). Thus, the model with the highest F1 did not provide the highest coverage. These results do not establish architectural superiority because calibration results and repeated runs are unavailable. A formalized confirmatory protocol is proposed with identity-disjoint data splits, separate calibration, model-validation, certification, and test subsets, five training seeds, temperature scaling, ECE, NLL, Brier score, AURC, and AUGRC. The certification-sample size is planned a priori from expected coverage, target risk, and the required probability of successful verification; a 400-item subset is treated only as a preliminary assessment. The cost model includes human review, erroneous automatic decisions, OPEX, and CAPEX. Under illustrative assumptions, the preferred model changes at an error cost of approximately RUB 4,262. The numerical findings remain pilot results; confirmatory claims require a full multi-seed experiment and an adequately sized independent certification sample.

Keywords: selective classification; coverage; selective risk; probability calibration; risk certification; AURC; AUGRC; human-in-the-loop; PIM; MDM; computer vision; AI economics.