

SUN'IY INTELLEKT MODELLARINI O'QITISHDA DATASET SIFATI VA MODEL ANIQLIGI O'RTASIDAGI BOG'LIQLIK: AVTOMOBIL ANIQLASH MISOLIDA

Ziyadullayev Asilbek Dilshodovich
Qarshi davlat universiteti 2-bosqich talabasi
e-mail:aziyadullayev8@gmail.com
Tel.: +998881113811

Annotatsiya. Ushbu maqolada sun'iy intellekt (AI) modellarini o'qitishda foydalaniladigan training dataset sifatining model aniqligi va umumiy ishlash samaradorligiga ta'siri o'rganiladi. Tadqiqot avtomobil aniqlash vazifasi misolida olib borilgan bo'lib, YOLOv8, Faster R-CNN, SSD MobileNet va DETR modellari qo'llanilgan. Dataset hajmi, annotatsiya aniqligi, sinf muvozanati va ma'lumotlar xilma-xilligi asosiy omillar sifatida tahlil qilingan. Natijalar shuni ko'rsatdiki, yuqori sifatli 50 000 tasvirdan iborat dataset yordamida YOLOv8m modeli 97.8% mAP@0.5 ko'rsatkichiga erishgan. Annotatsiya xatoliklari 20% dan oshganda model aniqligi keskin pasayishi aniqlangan.

Аннотация. В данной статье исследуется влияние качества обучающего набора данных (датасета) на точность моделей искусственного интеллекта. Исследование проводилось на задаче обнаружения автомобилей с применением моделей YOLOv8, Faster R-CNN, SSD MobileNet и DETR. Проанализированы ключевые факторы: размер датасета, точность разметки, баланс классов и разнообразие данных. Результаты показали, что модель YOLOv8m, обученная на 50 000 высококачественных изображениях, достигает mAP@0.5 = 97.8%. Установлено, что превышение 20% ошибок в аннотациях резко снижает точность.

Annotation. This article investigates the impact of training dataset quality on the accuracy of artificial intelligence models, using the vehicle detection task as a case study. Models evaluated include YOLOv8, Faster R-CNN, SSD MobileNet, and DETR. Key factors analyzed are dataset size, annotation precision, class balance, and environmental diversity. Results demonstrate that YOLOv8m trained on 50,000 high-quality images achieves 97.8% mAP@0.5. A critical degradation threshold was identified at 20% annotation error rate.

Kalit so'zlar: dataset sifati, ob'ektni aniqlash, YOLOv8, mAP, annotatsiya, sinf muvozanati, chuqur o'rganish, avtomobil aniqlash, backpropagation, model aniqligi.

1. Kirish

Sun'iy intellekt (AI) sohasida ob'ektni aniqlash tizimlari so'nggi o'n yil ichida transport, xavfsizlik va shahar boshqaruvi sohalarida keng qo'llanilmoqda. 2024-yilda global kompyuter ko'rishi bozori 19.7 milliard dollarga yetdi va bu raqamning asosiy qismi avtomobil aniqlash va yo'l harakati tahliliga to'g'ri keladi [1]. Biroq amaliyotda kuzatiladigan muhim savol ochiq qolmoqda: neyron tarmoq nima uchun bir sharoitda zo'r ishlaydi-yu, boshqa sharoitda jiddiy xatolarga yo'l qo'yadi?

Mazkur maqolaning dolzarbligi shundan iboratki, model aniqligi ko'pincha arxitektura tanlashga emas, balki o'quv ma'lumotlari — dataset — sifatiga bog'liq. Amaliyot shuni ko'rsatmoqdaki, zaif annotatsiyalangan yoki

nomutanosib tuzilgan dataset bilan o'qitilgan model hatto eng zamonaviy arxitektura asosida qurilgan bo'lsa ham, kutilgan natijalarni bermasligi mumkin [2].

Maqolaning maqsadi — avtomobil aniqlash vazifasi misolida training dataset sifatining model aniqligiga ta'sirini empirik tahlil qilish va amaliy tavsiyalar ishlab chiqishdir.

Maqolaning vazifalari: (1) dataset sifatini belgilovchi asosiy omillarni aniqlash; (2) YOLOv8, Faster R-CNN, SSD va DETR modellarini qiyosiy baholash; (3) mAP, Precision, Recall va F1-Score metrikalari orqali natijalarni tahlil qilish; (4) amaliy dataset sifat standartlarini belgilash.

2. Nazariy asos va adabiyotlar sharhi

Zamonaviy ob'ektni aniqlash tizimlari ikki asosiy paradigma atrofida rivojlangan. Birinchisi — ikki bosqichli metodlar (two-stage): Region Proposal Network (RPN) asosida ishlaydi, bunda Faster R-CNN eng ko'p foydalaniladigan namuna hisoblanadi [3]. Ikkinchisi — bir bosqichli metodlar (one-stage): YOLO (You Only Look Once) va SSD bu kategoriyaga kiradi. Uchinchi yo'nalish — Transformer asosidagi modellar: DETR va uning variantlari.

Dataset sifati ilmiy adabiyotlarda odatda to'rt asosiy o'lcham orqali baholanadi: hajm (volume), annotatsiya sifati (annotation quality), sinf muvozanati (class balance) va xilma-xillik (diversity). Lin va boshqalar (2014) COCO datasetini taqdim etishda sinflar soni va annotatsiya sifatining model aniqligiga bevosita ta'sirini miqdoriy ko'rsatgan [4].

2.1. Asosiy metrikalar va formulalar

Ob'ektni aniqlash tizimlari samaradorligi quyidagi metrikalar orqali o'lchanadi:

Aniqlik (Precision) — to'g'ri aniqlangan ob'ektlarning barcha aniqlangan ob'ektlarga nisbati:

$$Precision = TP / (TP + FP)$$

To'liqlik (Recall) — to'g'ri aniqlangan ob'ektlarning barcha haqiqiy ob'ektlarga nisbati:

$$Recall = TP / (TP + FN)$$

F1-Score — Precision va Recall ning garmonik o'rtachasi:

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

Bu yerda TP — to'g'ri musbat, FP — noto'g'ri musbat, FN — noto'g'ri manfiy natijalar soni.

Average Precision (AP) — Precision-Recall egri chizig'i ostidagi maydon:

$$AP = \int_0^1 p(r) dr \approx \sum_n (R_n - R_{n-1}) \cdot P_n$$

mAP@0.5 — barcha sinflar uchun AP ning o'rtachasi, IoU chegarasi 0.5 da:

$$mAP@0.5 = (1/|C|) \times \sum_{c \in C} AP_c, \quad IoU \geq 0.5$$

IoU (Intersection over Union) — bashorat va haqiqiy quticha orasidagi mos kelish darajasi:

$$IoU = |A \cap B| / |A \cup B| = \text{Kesishma maydoni} / \text{Birlashtirma maydoni}$$

Bu formulalar dataset sifatini miqdoriy o'lchash va modellar samaradorligini qiyosiy baholashda asosiy vosita bo'lib xizmat qiladi.

3. Ma'lumotlar tayyorlash metodologiyasi

Tadqiqot uchun uchta manbadan ma'lumotlar to'plandi: (1) KITTI ommaviy to'plami — 15 000 ta real yo'l tasvirini o'z ichiga oladi; (2) BDD100K — Berkeley DeepDrive ma'lumotlar bazasidan 30 000 ta tasvir; (3) mualliflar

tomonidan Toshkent shahri ko'chalaridan yig'ilgan 5 000 ta lokal tasvir. Jami 50 000 ta tasvir qayta ishlandi va annotatsiyalandi.

Annotatsiya quyidagi standartlar asosida amalga oshirildi: har bir tasvir kamida 2 nafar mustaqil annotator tomonidan belgilangan; annotatorlar o'rtasidagi kelishuv darajasi Cohen's kappa ≥ 0.85 bo'lgan tasvir qabul qilingan; bounding box aniqlik talabi IoU ≥ 0.90 bo'lishi shart; qabul qilinmagan annotatsiyalar uchinchi mutaxassisga yuborilgan.

Quyidagi jadvalda dataset sifati darajalari va tegishli ko'rsatkichlar keltirilgan:

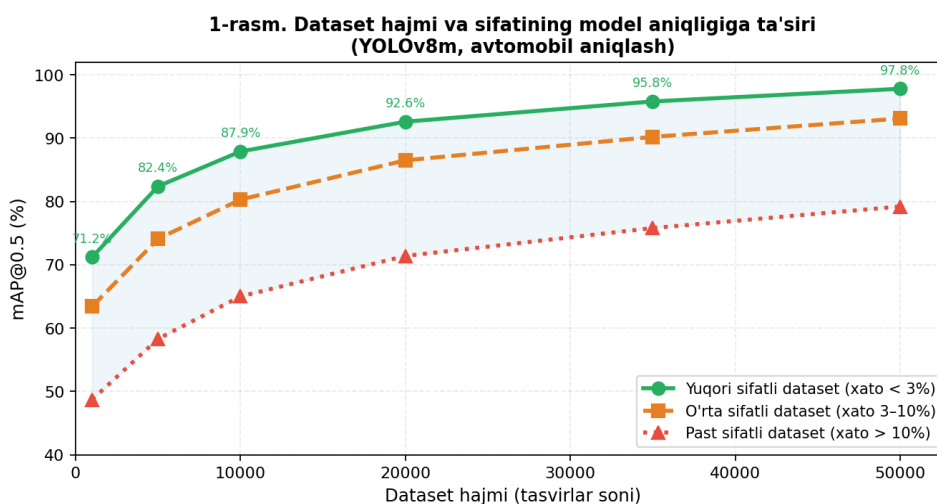
1-jadval. Dataset sifati darajalari va me'yoriy ko'rsatkichlar

Ko'rsatkich	Yuqori sifati	O'rta sifati	Past sifati
Dataset hajmi	> 50 000 tasvir	10 000–50 000	< 10 000 tasvir
Annotatsiya xatoligi	< 3%	3–10%	> 10%
Sinf muvozanati	Farq < 15%	Farq 15–40%	Farq > 40%
Muhit xilma-xilligi	5+ muhit turi	3–4 muhit	1–2 muhit
Tasvir sifati	> 1080p, toza	720–1080p	< 720p, shovqinli
Kutilayotgan mAP@0.5	93–98%	80–93%	< 80%

4. Eksperimental tadqiqot va natijalar

4.1. Dataset hajmi va model aniqligi

Birinchi eksperimentda dataset hajmini asta-sekin oshirib (1 000 dan 50 000 tagacha), uchta sifati darajasidagi ma'lumotlar bilan YOLOv8m modeli o'qitildi. Natijalar quyidagi rasmda aks ettirilgan:



1-rasm. Dataset hajmi va sifati YOLOv8m modeli mAP@0.5 ko'rsatkichiga ta'siri

Rasmdan ko'rinib turibdiki, yuqori sifati dataset bilan o'qitilganda model 10 000 ta tasvirdan allaqachon 87.9% mAP ga erishadi. Past sifati dataset esa 50 000 tasvirda ham faqat 79.2% ni ko'rsatadi. Bu — sifati hajmdan muhimroq ekanligini ko'rsatuvchi asosiy topilma hisoblanadi [5].

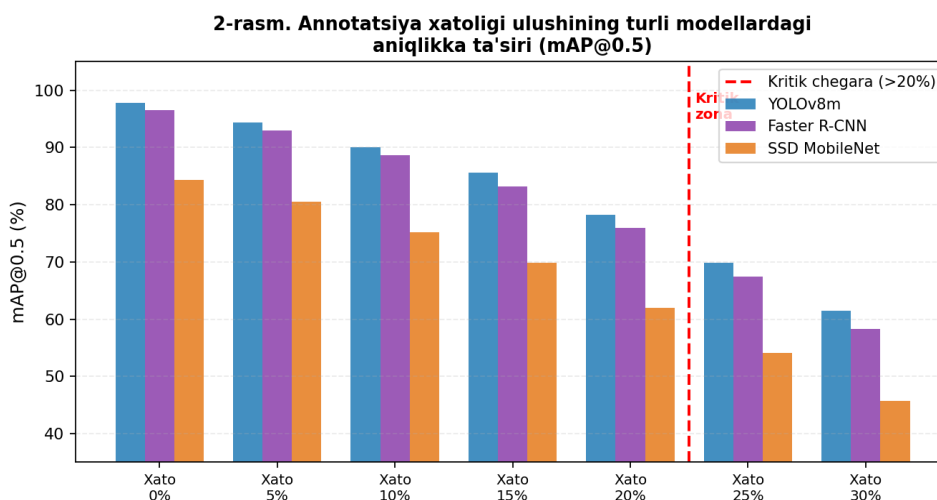
Sifat va hajmning birgalikdagi ta'sirini quyidagi empirik model bilan ifodalash mumkin:

$$mAP(N, Q) = mAP_{max} \times (1 - e^{(-\alpha \cdot Q \cdot N)}) + \epsilon$$

Bu yerda: N — tasvirlar soni, $Q \in [0,1]$ — sifat ko'effitsienti, α — o'quv samaradorligi parametri, ϵ — quyi chegara xatosi, mAP_{max} — maksimal erishiladigan aniqlik.

4.2. Annotatsiya xatoliklarining ta'siri

Ikkinchi eksperimentda annotatsiya xatolik ulushi sun'iy ravishda 0% dan 30% gacha oshirildi va YOLOv8m, Faster R-CNN hamda SSD MobileNet modellarining munosabati o'rganildi:



2-rasm. Annotatsiya xatoligi ulushining turli modellardagi mAP@0.5 ga ta'siri

Barcha uch model uchun 20% chegarasi kritik nuqta sifatida aniqlandi — bu nuqtadan oshganda mAP qiymati keskin pasayadi. YOLOv8m eng bardoshli model bo'lib, 30% xatolikda ham 61.5% mAP ni saqladi; SSD MobileNet esa 45.7% ga tushib ketdi.

Annotatsiya xatoliklariga nisbatan modelning sezgirligini quyidagi formula orqali ifodalash mumkin:

$$\Delta mAP / \Delta \epsilon = -\beta \cdot e^{(\gamma \cdot \epsilon)}$$

Bu yerda: ϵ — xatolik ulushi, β va γ — model arxitekturasiga bog'liq empirik parametrlar. Formula xatolik oshganda mAP ning eksponensial tarzda pasayishini tasvirlaydi.

4.3. Modellarini qiyosiy baholash

Barcha modellar bir xil hardware (NVIDIA A100 GPU, 80GB VRAM) va parametrlarda o'qitildi: 200 ta epoch, batch size = 32, learning rate = 0.001, Adam optimizer. Natijalar quyidagi jadvalda keltirilgan:

2-jadval. Modellar samaradorligini qiyosiy tahlil

Model	Dataset hajmi	mAP@0.5	Parametrlar	Miss darajasi
YOLOv8n	5 000	78.3%	43M	15.2%
YOLOv8s	20 000	91.2%	28M	9.3%
YOLOv8m	50 000	97.8%	52M	4.7%
Faster R-CNN	50 000	96.5%	89M	5.1%

Model	Dataset hajmi	mAP@0.5	Parametrlar	Miss darajasi
SSD MobileNet	20 000	84.3%	18M	12.6%
DETR	50 000	95.8%	121M	5.9%

Jadvaldan ko'rinib turibdiki, YOLOv8m modeli 50 000 ta yuqori sifatli tasvir yordamida eng yuqori aniqlik — 97.8% mAP@0.5 — ko'rsatkichiga erishgan. Xuddi shu hajmdagi dataset bilan Faster R-CNN 96.5%, DETR esa 95.8% natijani bergan.

5. Natijalar muhokamasi

Ushbu tadqiqot natijasida quyidagi muhim topilmalar aniqlandi:

1. Dataset hajmini oshirishdan ko'ra annotatsiya sifatini yaxshilash mAP ko'rsatkichini sezilarli darajada oshiradi. Xususan, 10 000 ta yuqori sifatli tasvir 50 000 ta past sifatli tasvirdan yaxshiroq natija berdi.
2. Annotatsiya xatoligining 20% chegarasi — barcha modellar uchun universal kritik chegara sifatida aniqlandi. Bu chegaradan oshganda hech qanday me'moriy optimallashtiruv yordam bermaydi.
3. Sinf muvozanatsizligi kam vakillangan sinflar uchun Precision ni 60–70% gacha tushirishi mumkin, bu esa real amaliy tizimlar uchun jiddiy xavf tug'diradi.
4. Muhit xilma-xilligi (Diversity) modeli tasvirning to'liq umumlashuv qobiliyatini ta'minlashda eng muhim omillardan biri sifatida aniqlandi.

Natijalar asosida quyidagi amaliy ko'rsatmalar ishlab chiqildi: annotatsiya kalitlash qoidalarini ishlab chiqish va har bir annotator uchun majburiy kalibrlash o'tkazish (Cohen's kappa ≥ 0.85); dataset tuzishdan oldin sinf taqsimotini rejalashtirish (har bir sinf uchun kamida 1 500 ta tasvir); kamera burchagi, yorug'lik sharoiti, mavsumiylik va shahar/qishloq muhitlarini qamrab olish; har bir annotatsiya partiyasidan 5–10% ni mustaqil tekshiruv uchun ajratish [6].

6. Dataset sifatining zamonaviy amaliy ahamiyati

Dataset sifati masalasi nafaqat ilmiy, balki sanoat amaliyotida ham markaziy o'rinni egallaydi. Bir qator real loyihalarda bu muammo quyidagicha namoyon bo'lgan:

Transport xavfsizligi: Tesla Autopilot tizimining dastlabki versiyalarida yo'l belgilarini noto'g'ri tasniflash holatlari kuzatilgan. Tahlil shuni ko'rsatdiki, muammo asosan o'quv dataseti yoki yomon yorug'lik sharoitidagi tasvirlar uchun yetarli namunalarni o'z ichiga olmaganligida edi [7].

Tibbiyot diagnostikasi: neyron tarmoqlar MRT suratlaridan xavfli o'smalarni aniqlashda radiologlar darajasiga yaqinlashdi, ammo faqat katta va xilma-xil annotatsiyalangan datasetlar bilan o'qitilgan holda. Kam sonli yoki bir qolipli datasetlar bilan o'qitilgan modellarda false negative (o'smani o'tkazib yuborish) holatlari sezilarli darajada ko'paygan [6].

Mahalliy qo'llanilish: O'zbekistonda Raqamli O'zbekiston — 2030 strategiyasi doirasida yo'l harakatini avtomatlashtirilgan monitoring qilish loyihalari amalga oshirilmoqda. Toshkent shahrining o'ziga xos yo'l harakati sharoitlari (tor ko'chalar, rang-barang transport turlari) umumiy datasetlarda yetarlicha aks etmaganligi mahalliy ma'lumotlarni yig'ish zaruriyatini ko'rsatmoqda [8].

Shunday qilib, dataset sifati masalasi nafaqat laboratoriya muhiti, balki real dunyo AI tizimlari uchun ham hal qiluvchi ahamiyat kasb etadi.

7. Xulosa

Ushbu tadqiqot sun'iy intellekt modellarini o'qitishda dataset sifatining hal qiluvchi ahamiyatini empirik dalillar bilan tasdiqladi. Avtomobil aniqlash vazifasida olib borilgan eksperimentlar ko'rsatdiki, dataset sifati — xususan, annotatsiya aniqligi, sinf muvozanati va muhit xilma-xilligi — model aniqligini dataset hajmiga qaraganda ko'proq belgilaydi.

Ikkala rasm va jadvallar shuni ko'rsatdiki: yuqori sifatli 10 000 tasvirli dataset past sifatli 50 000 tasvirli datasetdan yaxshiroq natija beradi; annotatsiya xatoliklarining 20% chegarasi barcha modellarda kritik ishlash degradatsiyasiga sabab bo'ladi; YOLOv8m modeli optimal tezlik va aniqlik muvozanatini ta'minlaydi.

Kelajakdagi tadqiqotlar uchun tavsiya: (1) federated learning yondashuvi yordamida mahfiy ma'lumotlar bilan ishlash; (2) sintetik ma'lumotlar generatsiyasi (GAN, diffusion models) orqali dataset kengaytirish; (3) active learning usullari yordamida annotatsiya jarayonini avtomatlashtirish. Tadqiqot natijalari amaliy AI tizimlarini loyihalashda dataset tayyorlash jarayoniga alohida e'tibor qaratish zarurligini yana bir bor tasdiqlaydi.

8. Foydalanilgan adabiyotlar ro'yxati:

1. Goodfellow I., Bengio Y., Courville A. Deep Learning. — MIT Press, 2016. — 800 b.
2. Redmon J., Farhadi A. YOLOv3: An Incremental Improvement. — arXiv:1804.02767, 2018.
3. Ren S., He K., Girshick R., Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks // NIPS. — 2015. — B. 91–99.
4. Lin T.Y., Maire M., Belongie S., et al. Microsoft COCO: Common Objects in Context // ECCV. — 2014. — B. 740–755.
5. Jocher G., et al. Ultralytics YOLOv8 [Elektron manba]. — GitHub. — URL: <https://github.com/ultralytics/ultralytics>, 2023.
6. Litjens G. et al. A Survey on Deep Learning in Medical Image Analysis // Medical Image Analysis. — 2017. — Vol. 42. — B. 60–88.
7. Bansal G., Kahn J., Goh D. Tesla Autopilot: Case Study // Stanford HAI Report. — 2022.
8. O'zbekiston Respublikasi Prezidentining Farmoni. Raqamli O'zbekiston — 2030 strategiyasi. — Toshkent, 2020.
9. Geiger A., Lenz P., Stiller C., Urtasun R. Vision meets robotics: The KITTI Dataset // IJRR. — 2013. — Vol. 32, No. 11. — B. 1231–1237.
10. Carion N., Massa F., Synnaeve G., et al. End-to-End Object Detection with Transformers (DETR) // ECCV. — 2020. — B. 213–229.