Maciej Stefańczyk
Instytut Automatyki i Informatyki Stosowanej, Politechnika Warszawska
ORCID: 0000-0001-9948-6319
e-mail: maciej.stefanczyk@pw.edu.pl

Wojciech Dudek
Instytut Automatyki i Informatyki Stosowanej, Politechnika Warszawska
ORCID: 0000-0001-5326-1034
e-mail: wojciech.dudek@pw.edu.pl

HYBRYDOWY KLASYFIKATOR TREŚCI CSAM W MATERIALE FOTOGRAFICZNYM[1]

Streszczenie

Poczucie anonimowości, połączone z łatwością publikowania treści w internecie, sprzyja pojawianiu się w sieci coraz większej ilości materiałów nielegalnych. Jedną z kategorii takich treści są materiały przedstawiające wykorzystywanie seksualne dzieci (Child Sexual Abuse Material – CSAM), treści erotyczne z ich udziałem czy też w ich obecności. Obecnie możliwości reakcji na tego typu treści (m.in. blokowanie domen) mają pojedyncze ośrodki na poziomie krajowym. Ze względu na charakter przeglądanych treści zadanie to jest bardzo obciążające psychicznie, a każda metoda prowadząca do zmniejszenia ekspozycji pracowników na tego typu obrazy jest na wagę złota. Z tego względu narodziła się idea projektu APAKT[2] (Automatyczne Przeszukiwanie, Analiza i Klasyfikacja Treści), który w sposób automatyczny analizuje przekazane dane i priorytetyzuje je tak, aby analityk możliwie szybko mógł ocenić, czy dana domena powinna zostać zablokowana czy nie. W niniejszym artykule przedstawione zostały prace (i związane z nimi specyficzne trudności) nad stworzenie klasyfikatora treści w materiale fotograficznym. W ich wyniku powstał hybrydowy klasyfikator, łączący współczesne osiągnięcia z dziedziny sztucznej inteligencji w rozpoznawaniu obrazów z klasycznymi metodami wspomagania decyzji.

Słowa kluczowe: CSAM, bezpieczeństwo cyberprzestrzeni, uczenie głębokie, rozpoznawanie obrazów, detekcja obiektów.

[1] Praca finansowana przez Narodowe Centrum Badań i Rozwoju w ramach projektu nr: CYBERSECIDENT/455132/ III/NCBR/2020.

[2] E. Niewiadomska-Szynkiewicz et al., System wspomagający wykrywanie treści wizualnych i tekstowych zagrażających bezpieczeństwu dzieci w cyberprzestrzeni, Cybersecurity and Law 2023, nr 2(10), s. 202-220.

WSTĘP

Klasyfikacja materiałów CSAM jest zadaniem bardzo obciążającym dla operatorów krajowych centrów monitorowania i reagowania na tego typu incydenty. W Polsce taką instytucją jest zespół Dyżurnet.pl[3], który działa na zasadzie reakcji na zgłoszenia nadchodzące od użytkowników globalnej sieci. Do zadań pracowników tego zespołu należy przeglądanie i analiza zgłoszonych treści (często setek zdjęć bądź nagrań wideo), dla których należy najpierw określić, czy w ogóle są nielegalne, a następnie podjąć odpowiednie działania. Do pewnego stopnia zadanie to może być wspomagane istniejącymi narzędziami wykrywającymi ponowne pojawienie się wcześniej sklasyfikowanych przez innych analityków obrazów (tzw. visual hash, np. PhotoDNA[4]), jednak rozwiązania te (choć często są dość zaawansowane i do pewnego stopnia zdolne do adaptacji takich jak przycięcie, modyfikacja pojedynczych elementów treści czy zdjęcia podobne do już istniejących), nie są w stanie efektywnie dopasować się do całkowicie nowych danych. Z tego względu zdecydowano się na podjęcie próby stworzenia systemu działającego w oparciu o metody uczenia maszynowego tak, aby podejmował decyzję na podstawie faktycznej treści obrazu a nie bezpośredniego porównania go z ograniczoną listą wstępnie sklasyfikowanych danych.

Przy tworzeniu systemu tego typu pojawiają się dodatkowe trudności i ograniczenia, nieobecne przy rozwijaniu typowych systemów rozpoznających. Kluczowym jest sama dostępność danych – oczywistym jest, że nie istnieją żadne publiczne, oetykietowane zbiory danych typu CSAM. Ze względów prawnych dane takie przechowywane mogą być legalnie jedynie wewnątrz zespołu Dyżurnet.pl, i nie ma do nich żadnego dostępu z zewnątrz. W ramach pracy w projekcie APAKT przygotowany został system umożliwiający testowanie algorytmów w bezpiecznym środowisku z dostępem do danych, jednak z punktu widzenia osób rozwijających algorytmy działał on jak czarna skrzynka – dostęp do wyników był ograniczony do suchych, zagregowanych do postaci tabelarycznej wyników, bez możliwości analizy i podglądu poszczególnych próbek uczących (szczególnie przydatnej przy analizie fałszywych odpowiedzi).

Naturalną odpowiedzią na ten problem wydaje się wytrenowanie systemów na publicznie dostępnych zbiorach danych legalnych (aczkolwiek często moralnie wątpliwych), takich jak pornografia dorosłych. Zbiory takie istnieją i pod pewnymi warunkami są udostępniane do badań[5], mogą też być potencjalnie stworzone przez agregację treści dostępnych w internecie. Może się wydawać, że łącząc tak wytrenowany system detekcji nagości z klasyfikatorem wieku można określić, czy zdjęcie przedstawia treści CSAM. Niestety, w rzeczywistości nie każde zdjęcie przedstawiające nagość i dziecko (a nawet nagie dziecko) jest materiałem typu CSAM, chociaż istnieje między

[3] Więcej zob. https://dyzurnet.pl/o-nas [dostęp: 13.01.2025].

[4] https://www.microsoft.com/en-us/photodna [dostęp: 13.01.2025].

[5] D. Moreira et al., Pornography classification: The hidden clues in video space–time, Forensic Science International 2016, nr 268, s. 46–61.

nimi pewna korelacja[6]. Definicja materiałów nielegalnych jest bardziej skomplikowana, opiera się między innymi na interakcjach osób widocznych na zdjęciu oraz tym, czy obraz skupia się na postaci dziecka. Z tego względu zdecydowano się na przygotowanie podejścia hybrydowego, w którym poszczególne aspekty składowe określane są na bazie ogólnodostępnych danych legalnych, a zestawy niezbędnych reguł są wyznaczane przez analizę wzajemnych relacji poszczególnych atrybutów obiektów wykrytych w obrazie. Taki system jest prostszy w wytworzeniu (do niektórych zadań można wykorzystać wstępnie wytrenowane modele), daje też pewne możliwości w kwestii wyjaśnialności zwracanych decyzji.

ZBIÓR DANYCH I MODELE ODNIESIENIE

Ze względu na opisane wcześniej problemy z dostępnością danych, dodatkową trudnością jest wykonanie analizy porównawczej z innymi opisywanymi w literaturze podejściami. Typowo wykorzystuje się do tego publiczne zbiory danych i uzyskiwane na nich metryki służą jako miara porównawcza. W tym przypadku zbiory takie nie istnieją, a w literaturze podawane są wyniki na prywatnych, niedostępnych zbiorach danych. Typowo autorzy podają jednak wyniki osiągane przez popularne modele jako punkt odniesienia, w związku z tym zastosowano takie samo podejście.

Do wytworzenia danych odniesienia (tzw. modelu „baseline") wykorzystano dwa popularne, dostępne wraz z wytrenowanymi wagami, modele NSFW (Not-Safe-For-Work), wytrenowane dla podobnego zadania detekcji treści niewskazanych do wyświetlania publicznego (głównie pornografii i nagości, ale też przemocy i materiałów drastycznych). Pierwszym wybranym modelem jest NudeNet[7], oparty na architekturze Xception[8]. Drugim jest OpenNSFW[9] (przygotowany przez pracowników Yahoo model oparty na architekturze ResNet-50[10]). Oba zwracają wynik w postaci oceny liczbowej danego obrazu, w skali od 0 (materiał całkowicie bezpieczny) do 1 (treść niebezpieczna).

Oba rozwiązania (a także system docelowy) trenowane i testowane były na zbiorze obrazów przygotowanym przez zespół ekspertów w ramach projektu APAKT. Zbiór wykorzystywany w trakcie tworzenia i testowania

[6] C. Laranjeira da Silva et al. Seeing without looking: Analysis pipeline for child sexual abuse datasets, [w:] Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 2022, s. 2189–2205.

[7] P. Bedapudi, NudeNet: An ensemble of neural nets for nudity detection and censoring, online: https://praneethbedapudi.medium.com/d9f3da721e3, 2019 [dostęp: 13.01.2025].

[8] F. Chollet, Xception: Deep learning with depthwise separable convolutions. [w:] Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, s. 1251-1258.

[9] J. Mahadeokar, G. Pesavento Open sourcing a deep learning solution for detecting NSFW images, online: https://yahooeng.tumblr.com/post/151148689421/ 2016. [dostęp: 13.01.2025].

[10] K. He et al., Deep residual learning for image recognition, [w:] Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, s. 770-778.

systemu składał się z ponad 30 tys. obrazów, podzielonych na dwie główne klasy (CSAM i nie-CSAM, o zbliżonej liczności), a każda z nich podzielona dodatkowo na podklasy (odpowiednio dla CSAM: „baseline" i „national" w zależności od wieku, oraz nie-CSAM: erotyka dziecięca, nagość dziecięca, pornografia dorosłych, i pozostałe). Należy tutaj zwrócić uwagę na obecność klasy „erotyka dziecięca" w kategorii nie-CSAM – co dodatkowo utrudnia całe zadanie i tym bardziej podkreśla trudność w prostym mapowaniu systemów detekcji nagości/pornografii na dziedzinę CSAM. Podział na poszczególne klasy i ich liczność przedstawione zostały w tabeli 1.

Klasa	**Liczność**	**Podklasa**	**Liczność**
CSAM	13 539	CSAM baseline	8785
		CSAM national	4754
nie-CSAM	16 574	Erotyka dziecięca	4259
		Nagość dziecięca	329
		Pornografia dorosłych	3779
		Pozostałe	8207

Tabela 1: Zestawienie liczności klas w zbiorze danych APAKT

W pierwszej kolejności skupiono się na uzyskaniu wyników wspomnianych wyżej modeli bazowych. Przeprowadzono trzy eksperymenty, o rosnącym stopniu dopasowania modeli do dziedziny CSAM. W pierwszym wykorzystano modele z dostarczonymi przez autorów wagami treningowymi do klasyfikacji danych w zbiorze CSAM przy prostym założeniu bezpośredniego mapowania wyjścia, tzn. przyjmując wyjście 0 za materiały nie-CSAM a 1 za CSAM. Zgodnie z oczekiwaniami, skuteczność takiego podejścia była słaba, a głównym źródłem błędów w tym wypadku było przypisywanie wysokich wyników zdjęciom nie-CSAM (zgodnie z tym, jak modele były wytrenowane, traktowały każdą pornografię tak samo jak CSAM). Pojawiały się też próbki CSAM mające bardzo niskie wyniki, jednak ze względu na brak wglądu w dane ciężko jest jednoznacznie stwierdzić, z czego mogło to wynikać. W odróżnieniu od typowych systemów klasyfikacji danych, gdzie do ewaluacji i porównania używa się miary F1 (ważona równomiernie czułość i precyzja), w systemach detekcji CSAM zwykle używana jest miara F2, gdzie czułość ma dwa razy większą wagę od precyzji. Oznacza to, że dużo większy wpływ na wynik ma błędne odrzucenie danych CSAM niż klasyfikacja nie-CSAM jako CSAM. Motywacja w tym przypadku jest taka, że ostatecznie i tak analityk musi ocenić dany materiał, więc bezpieczniej jest przepuścić kilka zdjęć legalnych niż odrzucić jedno nielegalne. Po zastosowaniu tej miary do odpowiedzi modeli uzyskano wynik 60% dla NudeNet i 73% dla OpenNSFW.

W drugim eksperymencie zachowano moduł ekstrakcji cech wybranych modeli, jednak zamiast klasyfikatora zastosowano model maszyny wektorów nośnych (SVM), wytrenowany na zbiorze APAKT. Zmiana ta spowodowała

wzrost osiąganych przez oba modele wyników, odpowiednio do 76% dla NudeNet i 79% dla OpenNSFW. Ostatni z eksperymentów polegał już na pełnym dopasowaniu modeli do nowych danych (trening całej sieci), i pozwolił na kolejne zwiększenie wyniku do odpowiednio 79% i 82%.

PODEJŚCIE HYBRYDOWE

Zgodnie z opisanymi wcześniej założeniami, zamiast stosowania jednego modelu głębokiego, zdecydowano się na zastosowanie hybrydy kilku algorytmów rozpoznających poszczególne elementy obrazu (z których każdy może być wytrenowany na dużym zestawie danych legalnych), oraz zbudowanie systemu decyzyjnego korzystającego z ich zagregowanych wyników. Detekcja osób na obrazie jest zadaniem dość dobrze zdefiniowanym i z dostępnymi wieloma modelami, które po dodatkowym treningu (wykonanym przez innych członków zespołu projektowego) udało się dostosować do wykrywania nawet sylwetek widocznych jedynie częściowo. Do detekcji intymnych części ciała oraz twarzy wykorzystano moduł detekcji wspomnianego wcześniej systemu NudeNet.

Te trzy elementy (sylwetka, intymne części ciała oraz twarz) stanowią pierwszy stopień analizy. Drugi etap agreguje i wzbogaca te dane o dodatkowe cechy. Jedną z cech, które odróżniają materiały typu CSAM od pozostałych jest celowe skupienie uwagi na postaci (bądź genitaliach) dziecka. Dla klasycznych zdjęć rodzaj ujęcia (szeroki kadr, zbliżenie) jest zwykle określany względem proporcji wielkości twarzy do wielkości kadru[11], jednak w wypadku materiałów CSAM twarz może nie być widoczna. W takiej sytuacji zaadaptowane zostało podejście bazujące na twarzy i dodatkowy rodzaj ujęcia określany jest na podstawie największej obecnej w obrazie proporcji widocznych genitaliów do wielkości kadru.

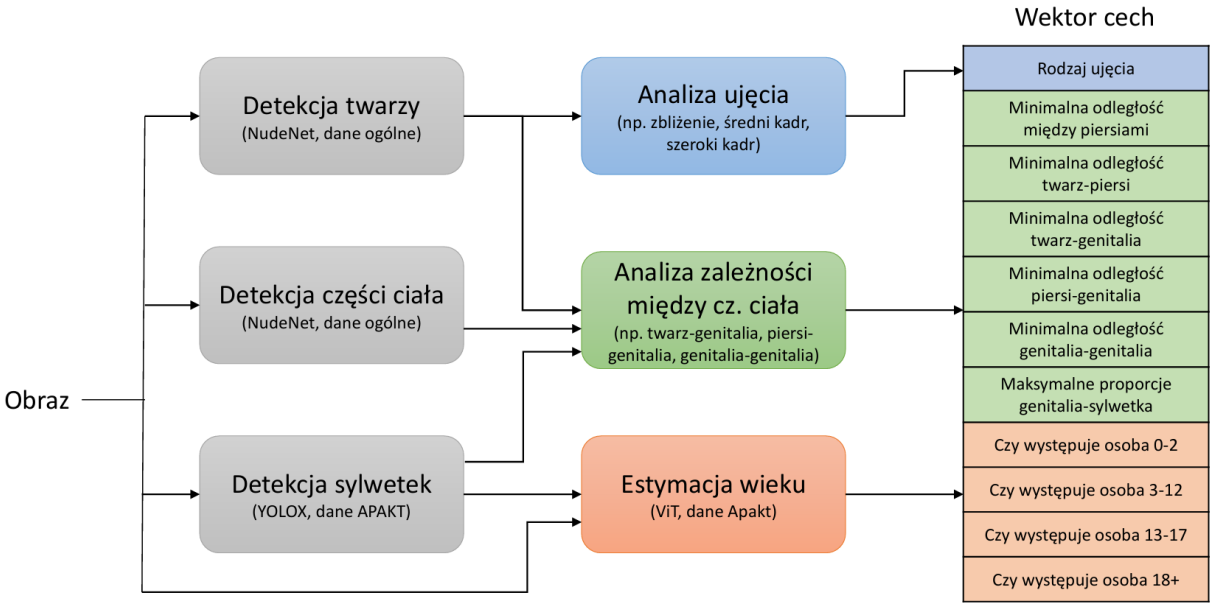
Relacje pomiędzy częściami ciała zostały ograniczone do znalezienia najbliższych do siebie par obiektów: piersi-piersi, twarz-piersi, genitalia-piersi, twarz-genitalia, genitalia-genitalia. Motywacją dla wyboru tego rodzaju cech jest opis potencjalnych interakcji między widocznymi na zdjęciu osobami. Zgodnie z przewidywaniami, największą rolę grały tutaj pary zawierające genitalia. Do estymacji wieku osób widocznych na zdjęciu opracowany w ramach projektu APAKT algorytm oparty o model Visual Transformer[12].

Ostateczny wektor cech dla systemu decyzyjnego zbudowany jest z połączenia 11 elementów: rodzaj ujęcia (jedna z siedmiu możliwych wartości), najbliższe odległości między wymienionymi wyżej parami obiektów (5 liczb z zakresu 0-1), największy widoczny obszar genitalny (liczba z zakresu 0-1) oraz cztery flagi określające, czy na zdjęciu widoczne są osoby z poszczególnych grup wiekowych. W przypadku, gdy któraś z wartości nie może zostać wyznaczona (np. brak widocznej twarzy), odpowiadający jej

[11] I. Cherif, V. Solachidis, I. Pitas, Shot type identification of movie content, [w:] 9th International Symposium on Signal Processing and Its Applications 2007, s. 1–4.

[12] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 2020.

element wektora cech oznaczany jest przez -1. Schematyczne przedstawienie przepływu danych oraz strukturę wektora cech przedstawiono na rysunku 1.



Rysunek 1: Struktura przepływu danych i wynikowy wektor cech

Ze względu na brak bezpośredniego dostępu do zbioru danych (oraz jego stosunkowo niewielki rozmiar) nie było możliwości stworzenia stałego podziału na zbiór treningowy i testowy. W zamian zastosowano podejście wielokrotnej walidacji skrośnej. Cały zbiór został podzielony na 5 równych części, po czym przeprowadzono 5 sesji treningowych, w których kolejno jedna z części była traktowana jako testowa a pozostałe jako treningowe. Ostateczny wynik powstał przez uśrednienie wyników poszczególnych podzadań.

Bazując na dość krótkim wektorze cech można było zastosować powszechnie znane algorytmy klasyfikacji danych. Zdecydowano się też na prowadzenie klasyfikacji wieloetapowej: wstępnego podziału na CSAM/nie-CSAM, po czym każda z kategorii była analizowana przez specjalizowany klasyfikator drugiego poziomu. Pierwszy poziom (potencjalnie najważniejszy z punktu widzenia projektu) uzyskał w testach wynik F2 na poziomie 89%. Podział na dwie podklasy CSAM osiągnął 88%, a na cztery klasy nie-CSAM jedynie 82% (najwięcej pomyłek następowało pomiędzy klasami dziecięcej nagości i erotyki).

W tym miejscu pojawia się kolejny problem z dostępnością do danych – nie ma prostej możliwości konfrontacji osiągniętych wyników z innymi opisywanymi w literaturze systemami. Z tego powodu często stosowane jest porównanie wyników osiąganych na swoim zbiorze testowym z wynikami pewnego modelu bazowego trenowanego na tym samym zbiorze. W tabeli 2 zawarto takie zestawienie dla opisywanego systemu oraz dwóch innych, opisanych w literaturze metod. Wszystkie rozwiązania stosowały sieć OpenNSFW jako model odniesienia, więc został on przedstawiony w tabeli. Ostateczną miarą skuteczności jest więc nie sam ostateczny wynik F2, a jego

przyrost względem modelu bazowego. Z kolei wynik modelu bazowego może świadczyć o samej jakości (trudności) posiadanego przez dany zespół zbioru danych. Im zbiór łatwiejszy, tym wyższy wynik był osiągany przez model OpenNSFW (w tabeli odnotowano także liczność zbioru treningowego i zawartych w nim zdjęć typu CSAM).

	F2	OpenNSFW	Przyrost	Zbiór danych
Prezentowany system	89,0	82,0	7,0	30 tys. (45% CSAM)
AttM-CNN[13]	93,2	86,8	6,4	2 mln. (0.3% CSAM)
2-Tiered SEIC[14]	87,7	80,7	7,0	59 tys. (57% CSAM)

Tabela 2: Porównanie wyników prezentowanego systemu z rozwiązaniami opisanymi w literaturze

PODSUMOWANIE

Zaprezentowany w artykule zestaw klasyfikatorów osiąga zadowalające wyniki dla zadania klasyfikacji binarnej CSAM/nie-CSAM. W zastosowaniu do wstępnej filtracji i kategoryzacji zgłoszeń może znacznie przyspieszyć pracę analityka, kierując jego uwagę na początku na najbardziej podejrzane obrazy, zwiększając możliwą do przeanalizowania liczbę domen przy jednoczesnym ograniczeniu do niezbędnego minimum kontaktu z materiałem wrażliwym. Pomimo, iż przygotowany system nie osiąga wyników istotnie lepszych niż pojedyncze, głębokie modele neuronowe, jego przewagą jest możliwość wyjaśnienia do pewnego stopnia podjętych decyzji (wskazanie konkretnych cech, które najbardziej wpłynęły na wynik). Dodatkowo adaptacja do nowych danych jest dużo szybsza niż dla modeli głębokich – zamiast aktualizacji wag całych modeli wystarczy przeprowadzić trening trzech niewielkich klasyfikatorów wynikowych.

Bibliografia

Bedapudi P., NudeNet: An ensemble of neural nets for nudity detection and censoring, online: https://praneethbedapudi.medium.com/d9f3da721e3, 2019. [dostęp: 13.01.2025]

Cherif I., V. Solachidis, I. Pitas, Shot type identification of movie content, [w:] 9th International Symposium on Signal Processing and Its Applications, 2007.

[13] A. Gangwar et al., AttM-CNN: Attention and metric learning based CNN for pornography, age and Child Sexual Abuse (CSA) detection in images, Neurocomputing 2021, nr 445, s. 81–104.

[14] P. Vitorino et al., Leveraging deep neural networks to fight child pornography in the age of social media, Journal of Visual Communication and Image Representation 2018, nr 50, s. 303–313.

Chollet, Xception F., Deep learning with depthwise separable convolutions, [w:] Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017.
Dosovitskiy A. et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2020.
Gangwar A. et al., AttM-CNN: Attention and metric learning based CNN for pornography, age and Child Sexual Abuse (CSA) detection in images, Neurocomputing, 2021, nr 445.
He K. et al., Deep residual learning for image recognition, [w:] Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
Laranjeira da Silva C. et al., Seeing without looking: Analysis pipeline for child sexual abuse datasets, [w:] Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022.
Mahadeokar J., Pesavento G., Open sourcing a deep learning solution for detecting NSFW images., online: https://yahooeng.tumblr.com/post/ 151148689421/, 2016. [dostęp: 13.01.2025]
Moreira D. et al., Pornography classification: The hidden clues in video space–time, Forensic Science International, 2016, nr 268.
Niewiadomska-Szynkiewicz E. et al., System wspomagający wykrywanie
treści wizualnych i tekstowych zagrażających bezpieczeństwu dzieci w cyberprzestrzeni, Cybersecurity and Law, 2023, nr 2(10).
Vitorino P. et al., Leveraging deep neural networks to fight child pornography in the age of social media, Journal of Visual Communication and Image Representation, 2018, nr 50.

HYBRID CSAM CONTENT CLASSIFIER IN PHOTOGRAPHIC MATERIAL

Abstract

A sense of anonymity, combined with the ease of publishing content on the Internet, has fostered the appearance of an increasing amount of illegal material online. One category of such content is Child Sexual Abuse Material (CSAM), erotic content with or in the presence of children. Currently, the ability to respond to this type of content (including domain blocking) is available to individual units at the national level. Due to the nature of the content, this task is mentally taxing, and any method leading to a reduction in employees' exposure to such images is at a premium. For this reason, the idea of the APAKT (Automated Content Search, Analysis and Classification) project was born, which automatically analyzes the submitted data and prioritizes it so that the analyst can assess as quickly as possible whether a domain should be blocked or not. This article presents the work (and related specific difficulties) on the creation of a content classifier in photographic material. As a result, a hybrid classifier was created, combining modern developments in artificial intelligence in image recognition with classical decision support methods.

Keywords: CSAM, cybersecurity, deep learning, image recognition, object detection