AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

DZIEDZINA: Nauki inżynieryjno-techniczne

DYSCYPLINA: Automatyka, Elektronika, Elektrotechnika i Technologie Kosmiczne

ROZPRAWA DOKTORSKA

Elektroniczne Systemy Zabezpieczenia Technicznego jako źródło danych w procesie automatycznej obrony przed cyberatakami oraz wykrywania sprawcy przestępstwa

Autor: Jan Kapusta

Promotor rozprawy: Prof. dr hab. inż. Jerzy Baranowski
Promotor pomocniczy: Dr inż. Waldemar Bauer

Praca wykonana: AGH, Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

Kraków, 2025

AGH

AGH UNIVERSITY OF KRAKOW

FIELD OF SCIENCE: Engineering and Technical Sciences

SCIENTIFIC DISCIPLINE: Automation, Electronics, Electrical Engineering and Space Technologies

DOCTORAL THESIS

Electronic Technical Security Systems as a Data Source in the Process of Automatic Defence Against Cyberattacks and Detection of Crime Perpetrators

Author: Jan Kapusta

First supervisor: Prof. dr hab. inż. Jerzy Baranowski
Assisting supervisor: Dr inż. Waldemar Bauer

Completed in: AGH University of Krakow, Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering

Krakow, 2025

Streszczenie

Uzasadnienie wyboru tematu i powiązanie z programem Doktorat Wdrożeniowy

Niniejsza rozprawa została zrealizowana w ramach programu *Doktorat Wdrożeniowy*, projekt pt. „Elektroniczne Systemy Zabezpieczenia Technicznego jako źródło danych w procesie automatycznej obrony przed cyberatakami oraz wykrywania sprawcy przestępstwa", co implikuje tytuł rozprawy.

Niniejsza rozprawa koncentruje się na analizie i modelowaniu *Cyber-Fizycznych Systemów Ochrony (CPPS)* z wykorzystaniem metod sztucznej inteligencji, co bezpośrednio odpowiada na problemy sformułowane w temacie badawczym. Takie podejście umożliwia z jednej strony ocenę efektywności zainstalowanych systemów ochrony, poprzez wykorzystanie danych pochodzących nie tylko z samych systemów zabezpieczeń, ale również przez uwzględnienie kontekstu bezpieczeństwa i estymowanie atrybutów potencjalnego adwersarza (np. jego motywacji czy kwalifikacji). Z drugiej strony, proponowana metodologia pozwala na planowanie racjonalnego, *evidence-based* wzmocnienia środków mitygacji zagrożeń pochodzących zarówno ze sfery fizycznej, jak i cyberprzestrzeni.

W ten sposób ESZT stają się nie tylko elementem infrastruktury ochronnej, ale przede wszystkim bogatym źródłem danych, które po odpowiedniej analizie i modelowaniu stanowią fundament dla zintegrowanych strategii cyber-fizycznej obrony.

Zidentyfikowane luki badawcze

Analiza stanu wiedzy ujawniła pięć kluczowych luk:

Po pierwsze, tradycyjna metodologia EASI (Estimated Adversary Sequence Interruption), wprowadzona w 1977 roku, nie uwzględnia zagrożeń cybernetycznych. Badania własne wykazały, że klasyczne podejście wskazywało 89,88% prawdopodobieństwo przerwania ataku, podczas gdy po uwzględnieniu cyberzagrożeń wartość ta spadła do 1,63%. **Po drugie**, brakuje modeli degradacji specyficznych dla elektronicznych systemów zabezpieczeń technicznych (ESZT). Opracowany probabilistyczny model degradacji uwzględnia pięć czynników: starzenie komponentów (rozkład normalny), wzrost wiedzy przeciwnika (rozkład normalny), postęp technologiczny narzędzi ataku (model logarytmiczny), pojawianie się podatności (proces Bernoulliego) oraz efekty konserwacji (rozkład normalny). Symulacje Monte Carlo wykazały, że efektywność systemu może spaść z 98% do około 25% w ciągu 10 lat, a średnia linia trendu przekracza poziom 50% około 75-80 miesiąca eksploatacji. **Po trzecie**, studium przypadku pokazało praktyczne konsekwencje ignorowania degradacji: klasyczna ocena EASI wskazywała 88,64% skuteczności, podczas gdy po uwzględnieniu degradacji wartość spadła do 76,19% – różnica 12,45 punktów

procentowych. **Po czwarte**, brak standaryzacji w reprezentacji elementów infrastruktury oraz automatyzacji detekcji symboli technicznych na planach. Konwencjonalne modele detekcji obiektów nie radzą sobie z interpretacją kontekstu, gdzie znaczenie symbolu zależy od otoczenia. **Po piąte**, istniejące rozwiązania są fragmentaryczne i adresują pojedyncze aspekty problemu, brakuje holistycznego podejścia integrującego wszystkie wymiary oceny CPPS.

Teza i metodologia

Teza pracy: możliwe jest opracowanie zintegrowanego, probabilistycznego modelu oceny skuteczności CPPS uwzględniającego jednocześnie degradację komponentów, zagrożenia cybernetyczne oraz kontekst operacyjny.

Proponowane rozwiązanie składa się z trzech warstw: (1) percepcyjnej/detekcji - kontekstowa analiza planów z detekcją obiektów COG (Contextual Object Grouping), (2) decyzyjnej - grafy ścieżek ataku z algorytmem A* i modyfikacją EASI o wskaźniki degradacji oraz element cyberodporności, (3) raportującej - automatyzacja z użyciem modeli językowych LLM.

Wyniki i wkład naukowy

Symulacje Monte Carlo wykazały, że uwzględnienie degradacji i cyberzagrożeń może znacząco obniżyć skuteczność systemu bezpieczeństwa. Tego rodzaju przeszacowanie efektywności systemu może nie tylko oznaczać brak gospodarności, ale prowadzić do istotnych konsekwencji dla infrastruktury krytycznej w przypadku skutecznego ataku.

Główny wkład stanowią: (1) stworzenie pierwszego probabilistycznego modelu degradacji ESZT w kontekście CPPS, (2) rozszerzenie EASI o komponenty cyber-odporności i degradację komponentów, (3) nowe podejście (COG) do kontekstowej interpretacji symboli technicznych, (4) stworzenie praktycznego narzędzia do dynamicznego zarządzania bezpieczeństwem. Praca integruje metody z dziedzin oceny ryzyka, sztucznej inteligencji i cyberbezpieczeństwa w spójną metodologię wspartą walidacją na rzeczywistych przypadkach.

Implementacja

Rezultatem pracy jest funkcjonalne narzędzie AI-SecPA implementujące pełny proces oceny, począwszy od detekcji elementów architektonicznych przez uczenie maszynowe po generowanie rekomendacji z wykorzystaniem LLM. Narzędzie wydaje się wprowadzać istotne wsparcie w procesie decyzyjnym z zakresu zarządzania bezpieczeństwem operacyjnym i cyber-fizycznym.

Abstract

Topic Selection and Connection to the Implementation Doctorate Program

This dissertation was carried out within the framework of the *Implementation Doctorate* program, project titled "Electronic Technical Security Systems as a Data Source in the Process of Automatic Defense Against Cyberattacks and Detection of Crime Perpetrators,"which implies the title of the dissertation.

This dissertation focuses on the analysis and modeling of *Cyber-Physical Protection Systems (CPPS)* using artificial intelligence methods, which directly addresses the problems formulated in the research topic. This approach enables, on one hand, the assessment of the effectiveness of installed protection systems by utilizing data not only from the security systems themselves but also by taking into account the security context and estimating attributes of potential adversaries (e.g., their motivation or qualifications). On the other hand, the proposed methodology allows for planning rational, *evidence-based* strengthening of threat mitigation measures originating from both the physical sphere and cyberspace.

In this way, Electronic Security Systems (ESS) become not only an element of protective infrastructure but, above all, a rich source of data that, after appropriate analysis and modeling, forms the foundation for integrated cyber-physical defense strategies.

Identified Research Gaps

Analysis of the state of knowledge revealed five key gaps:

First, the traditional EASI (Estimated Adversary Sequence Interruption) methodology, introduced in 1977, does not account for cyber threats. Our own research showed that the classical approach indicated an 89.88% probability of attack interruption, while after accounting for cyber threats, this value dropped to 1.63%. **Second**, there is a lack of degradation models specific to Electronic Security Systems (ESS). The developed probabilistic degradation model accounts for five factors: component aging (normal distribution), adversary knowledge growth (normal distribution), technological advancement of attack tools (logarithmic model), emergence of vulnerabilities (Bernoulli process), and maintenance effects (normal distribution). Monte Carlo simulations showed that system effectiveness can drop from 98% to approximately 25% over 10 years, with the average trend line crossing the 50% level around months 75-80 of operation. **Third**, a case study demonstrated the practical consequences of ignoring degradation: classical EASI assessment indicated 88.64% effectiveness, while after accounting for degradation, the value dropped to 76.19% – a difference of 12.45 percentage points. **Fourth**, there is a lack of standardization in infrastructure element representation and automation of technical symbol detection on plans. Conven-

tional object detection models cannot handle context interpretation, where symbol meaning depends on surroundings. **Fifth**, existing solutions are fragmentary and address individual aspects of the problem; there is a lack of holistic approaches integrating all dimensions of CPPS assessment.

Thesis and Methodology

Thesis: it is possible to develop an integrated, probabilistic model for assessing CPPS effectiveness that simultaneously accounts for component degradation, cyber threats, and operational context.

The proposed solution consists of three layers: (1) perceptual/detection – contextual plan analysis with COG (Contextual Object Grouping) object detection, (2) decision-making – attack path graphs with A* algorithm and EASI modification incorporating degradation indices and cyber-resilience element, (3) reporting – automation using LLM language models.

Results and Scientific Contribution

Monte Carlo simulations demonstrated that accounting for degradation and cyber threats can significantly reduce security system effectiveness. This type of system effectiveness overestimation may not only indicate lack of economy but lead to significant consequences for critical infrastructure in case of a successful attack.

Main contributions include: (1) creation of the first probabilistic degradation model for ESS in the CPPS context, (2) extension of EASI with cyber-resilience components and component degradation, (3) new approach (COG) for contextual interpretation of technical symbols, (4) creation of a practical tool for dynamic security management. The work integrates methods from risk assessment, artificial intelligence, and cybersecurity domains into a coherent methodology supported by validation on real-world cases.

Implementation

The result of this work is a functional AI-SecPA tool implementing the complete assessment process, from detection of architectural elements through machine learning to recommendation generation using LLM. The tool appears to provide significant support in decision-making processes regarding operational and cyber-physical security management.

Podziękowania

Podziękowania

Serdecznie dziękuję prof. dr hab. inż. Jerzemu Baranowskiemu, który jako jeden z naprawdę nielicznych uwierzył we mnie i sens tego co robię. Bez jego wiary, aktywnego wsparcia oraz motywowania, przejście przez cały proces uzyskiwania stopnia doktora nie byłoby możliwe.

.

Wielkie podziękowania należą się mojemu bezpośredniemu przełożonemu Steve'owi Casey, który nie tylko motywował mnie merytorycznie, ale także wspierał w zachowaniu odpowiedniego balansu między życiem zawodowym, pracą naukową oraz życiem prywatnym.

.

Mojemu promotorowi pomocniczemu dr inż. Waldemarowi Bauerowi, który wspierał mnie nie tylko tłumacząc elementy naukowo-techniczne, które musiałem opanować, ale również bez wahania poprawiał i szlifował treści moich artykułów.

.

Wielkie podziękowania należą się również całemu zespołowi Laboratorium Informatyki w Sterowaniu i Zarządzaniu Katedry Automatyki i Robotyki AGH, którego czuję się częścią w ramach moich prac badawczych. To oni budują fantastyczną atmosferę sprzyjającą współpracy. ·

Dziękuję żonie, że mnie namówiła, by kontynuować edukację i dzielnie znosi konsekwencje tej decyzji.

.

Wreszcie podziękowania należą się mojemu pracodawcy CISCO Systems Poland sp. z o.o. oraz Ministerstwu Nauki i Szkolnictwa Wyższego, dzięki któremu mogłem realizować swój projekt badawczy w ramach programu "Doktorat Wdrożeniowy".

Spis treści

1. Wprowadzenie ... 16

1.1. Uzasadnienie wyboru tematu i powiązanie z programem Doktorat Wdrożeniowy ... 16

1.2. Kontekst problemu ... 16

1.3. Główne wyzwania i luki ... 17

1.4. Konsekwencje obecnego stanu ... 26

1.5. Identyfikacja luk badawczych ... 27

1.6. Cel i zakres pracy ... 28

1.7. Teza pracy ... 30

1.8. Struktura rozprawy ... 30

1.9. Podsumowanie ... 30

2. Zasady oceny systemów bezpieczeństwa fizycznego i zidentyfikowane deficyty ... 32

2.1. Metodologia EASI - przypomnienie ... 32

2.2. Zidentyfikowane deficyty klasycznych podejść ... 33

2.3. Model probabilistyczny degradacji systemów elektronicznych ... 43

2.4. Łączny wpływ cyberzagrożeń i degradacji ... 48

2.5. Podsumowanie ... 49

3. Detekcja elementów budynku i infrastruktury bezpieczeństwa ... 50

3.1. Ewolucja detekcji obiektów w kontekście bezpieczeństwa ... 50

3.2. Architektura YOLO — ewolucja do YOLOv8 ... 51

3.3. Konfiguracja eksperymentalna ... 53

3.4. Strategia augmentacji danych ... 55

3.5. Post–processing i rafinacja detekcji ... 55

3.6. Wyniki eksperymentalne i analiza ... 56

3.7. Detekcja elementów systemów bezpieczeństwa ... 57

3.8. Rozważania implementacyjne ... 61

3.9. Kierunki dalszego rozwoju ... 62

3.10. Podsumowanie ... 65

4. Konstruowanie Grafów Obiektów i Heurystyczna Ocena Ścieżek Ataku z Wykorzystaniem Algorytmu A* ... 66
4.1. Wprowadzenie teoretyczne do teorii grafów ... 66
4.2. Algorytmy przeszukiwania grafów - ewolucja i zastosowania ... 70
4.3. Algorytm A* - teoria i zaawansowane zastosowania ... 71
4.4. Funkcje heurystyczne w analizie bezpieczeństwa ... 72
4.5. Teoria gier w modelowaniu przeciwnika ... 74
4.6. Integracja z systemem COG - teoretyczne podstawy ... 75
4.7. Eksploracja wielościeżkowa - podstawy teoretyczne ... 76
4.8. Walidacja i filtracja ścieżek ... 80
4.9. Integracja z metodologią EASI ... 82
4.10. Zastosowania w cyber-fizycznych systemach bezpieczeństwa ... 85
4.11. Podsumowanie teoretyczne ... 88
5. Zastosowanie dużych modeli językowych w konsolidacji i analizie danych bezpieczeństwa ... 91
5.1. Duże modele językowe - podstawy teoretyczne ... 91
5.2. Teoretyczne podstawy integracji LLM w automatyzacji oceny systemów bezpieczeństwa.. 94
5.3. Teoretyczne podstawy strukturyzowanej inżynierii promptów w domenie bezpieczeństwa. 95
5.4. Wykrywanie wzorców i anomalii jako właściwość LLM ... 97
5.5. Konsolidacja i strukturyzacja danych dla potrzeb analitycznych ... 99
5.6. Zarządzanie tokenizacją i ekonomia LLM w aplikacjach bezpieczeństwa ... 102
5.7. Synteza teoretyczna i implikacje metodologiczne ... 103
6. Implementacja narzędzia: AI-SecPA ... 107
6.1. Część I: Przygotowanie danych i kontekstu ... 107
6.2. Część II: Detekcja i analiza obrazu ... 111
6.3. Część III: Modelowanie grafowe i analiza ścieżek ataku ... 115
6.4. Część IV: Synteza wyników z wykorzystaniem LLM ... 130
7. Konkluzje i analiza krytyczna ... 135
7.1. Główne osiągnięcia badawcze ... 135
7.2. Analiza krytyczna ograniczeń ... 136
7.3. Wkład w stan wiedzy ... 138
7.4. Kierunki dalszych badań ... 139
7.5. Podsumowanie ... 140
A. Dodatek A - Raport LLM ... 142

Glosariusz

Z uwagi na implementacyjny charakter niniejszej rozprawy, w glosariuszu obok terminów o charakterze akademickim i technologicznym uwzględniono również pojęcia branżowe charakterystyczne dla obszaru bezpieczeństwa fizycznego oraz elektronicznych systemów zabezpieczeń. Celem takiego podejścia jest zapewnienie spójności terminologicznej oraz ułatwienie interpretacji fragmentów o charakterze praktycznym, gdzie język techniczny i specjalistyczny stanowi istotny element opisu. Wszystkie pojęcia zostały dobrane w sposób odzwierciedlający zarówno kontekst badawczy, jak i uwarunkowania implementacyjne pracy.

- **Infrastruktura krytyczna** — Zespół powiązanych systemów, sieci oraz aktywów fizycznych i wirtualnych, których awaria lub zniszczenie może poważnie naruszyć bezpieczeństwo państwa, zdrowie publiczne, gospodarkę lub funkcjonowanie społeczeństwa. Jest priorytetowym obszarem ochrony ze względu na wysokie skutki potencjalnych zakłóceń.

- **Atak** — Celowa sekwencja działań prowadząca do naruszenia poufności, integralności lub dostępności chronionych zasobów. Może mieć charakter fizyczny, cybernetyczny albo cyber-fizyczny (łączony).

- **Wektor ataku** — Kanał lub sposób, którym przeciwnik inicjuje i rozwija oddziaływanie na system (np. naruszenie perymetru, zdalne włamanie do sieci, manipulacja sterowaniem). Dobór wektora zależy od podatności celu i ryzyka detekcji.

- **Przeciwnik (adwersarz)** — Podmiot (osoba, grupa przestępcza, organizacja, aktor państwowy) planujący i realizujący działania wymierzone w system ochrony. Zdolności, zasoby i wiedza przeciwnika kształtują prawdopodobieństwo sukcesu ataku i wybór ścieżek.

- **Cyber-Fizyczne Systemy Ochrony (CPPS)** — Zintegrowane systemy łączące środki techniczne, informatyczne i architektoniczne z procedurami organizacyjnymi oraz ochroną osobową. Celem jest przeciwdziałanie zagrożeniom zamierzonym i przypadkowym w środowisku IT/OT powiązanym z infrastrukturą fizyczną.

- **Systemy Ochrony Fizycznej (PPS)** — Podzbiór CPPS skupiony na barierach fizycznych, detekcji i reakcji na wtargnięcia. Klasyczne oceny PPS bazują m.in. na analizie warstw i prawdopodobieństwie przerwania sekwencji ataku.

- **Elektroniczne Systemy Bezpieczeństwa (ESS)** — Komponenty elektroniczne w PPS/CPPS (SSW, CCTV, kontrola dostępu, czujniki, komunikacja alarmu). Podlegają degradacji w czasie, co wpływa na realną skuteczność ochrony.

- **Zagrożenia hybrydowe** — Skorelowane działania łączące środki cybernetyczne, dezinformacyjne i kinetyczne. Siła oddziaływania wynika z jednoczesnego wpływu na wiele warstw systemu i jego otoczenia.

- **Atak cyber-fizyczny (blended attack)** — Działanie, w którym elementy cyber (np. przejęcie systemu sterowania) są sprzężone z elementami fizycznymi (np. sabotaż, wybuch). Wymaga łącznej oceny bezpieczeństwa IT/OT i ochrony fizycznej.

- **Sekwencja działań przeciwnika (ASS)** — Uporządkowany ciąg akcji od inicjacji po osiągnięcie celu. Modelowanie ASS pozwala wskazać punkty detekcji i reakcji oraz szacować prawdopodobieństwo przerwania.

- **Ścieżka ataku** — Konkretny wariant przejścia przez kolejne elementy i warstwy systemu w drodze do celu. W analizie grafowej reprezentowana jako ciąg węzłów i krawędzi od punktu wejścia do zasobu o wysokiej wartości.

- **Warstwa ochronna** — Funkcjonalny poziom zabezpieczeń (np. perymetr, nadzór wizyjny, kontrola dostępu, czujniki wewnętrzne). Sumaryczny efekt wielu warstw zwiększa szanse detekcji i skutecznej reakcji.

- **Ochrona obwodowa (perymetr)** — Pierwsza linia obrony wyznaczająca granicę między strefą kontrolowaną a niekontrolowaną (np. ogrodzenie z czujnikiem światłowodowym). Jej naruszenie zwykle inicjuje łańcuch zdarzeń detekcyjnych.

- **Nadzór wizyjny (CCTV)** — System kamer i analityki obrazu wspierający detekcję, weryfikację alarmów i śledzenie zdarzeń. Skuteczność zależy od rozmieszczenia, jakości obrazu i procedur reakcji.

- **Kontrola dostępu** — Mechanizmy ograniczające wejście do stref/obiektów (czytniki, zamki, kontrolery) sprzężone z logiką uprawnień i rejestracją zdarzeń. Dostarczają punktów detekcji/odrzucenia w modelu oceny.

- **System Sygnalizacji Włamania (SSW)** — Sensory (np. PIR, kontaktrony) i centrala alarmowa odpowiedzialne za wykrycie, klasyfikację i sygnalizację naruszeń. Integracja z komunikacją i reakcją jest krytyczna dla całkowitej skuteczności.

- **EASI (Estimated Adversary Sequence Interruption)** — Metodyka szacowania prawdopodobieństwa przerwania sekwencji ataku przed osiągnięciem celu. Operuje na miarach detekcji, komunikacji alarmu i skuteczności reakcji w kolejnych krokach ASS.

- **Prawdopodobieństwo przerwania ataku,** $P(I)$ — Kluczowa miara w EASI określająca szansę zatrzymania przeciwnika zanim dotrze do celu. Jest funkcją prawdopodobieństw detekcji, komunikacji i reakcji wzdłuż sekwencji.

- **Prawdopodobieństwo detekcji** (P_{D_i}) — Szansa, że warstwa/element wykryje działanie przeciwnika na i-tym etapie. Wpływa bezpośrednio na $P(I)$ oraz na opłacalność poszczególnych ścieżek z perspektywy adwersarza.

- **Prawdopodobieństwo skutecznej komunikacji alarmu** (P_{C_i}) — Szansa, że informacja o detekcji dotrze do podmiotów reagujących w wymaganym czasie i jakości. Zakłócenia łączności obniżają realną efektywność systemu.

- **Warunkowe prawdopodobieństwo reakcji** ($P(R|A_i)$) — Szansa skutecznej interwencji po wykryciu akcji A_i, zależna m.in. od czasu dojazdu, gotowości i procedur. Zamyka łańcuch „detekcja–komunikacja–reakcja".

- **Efektywność CPPS wobec ataków cyber-fizycznych** (PE_{cp}) — Miara łączna uwzględniająca komponenty cyber i fizyczny. Modelowana jako komplementarne połączenie prawdopodobieństw sukcesu ataku cyber (PE_c) i fizycznego (PE_{pz}).

- **Degradacja (w bezpieczeństwie)** — Czasowy spadek skuteczności elementów ochrony spowodowany starzeniem, wzrostem kompetencji przeciwnika, rozwojem narzędzi ataku, pojawianiem się podatności oraz ograniczoną skutecznością utrzymania. Jej nieuwzględnienie prowadzi do przeszacowania $P(I)$ i błędnych decyzji inwestycyjnych.

- **Cyfrowy bliźniak (Digital Twin) w ochronie** — Wirtualny model instalacji sprzężony z danymi operacyjnymi, wykorzystywany do analizy ryzyka i testowania scenariuszy bez wpływu na produkcję. Umożliwia ocenę reakcji CPPS na zdarzenia cyber-fizyczne.

1. Wprowadzenie

1.1. Uzasadnienie wyboru tematu i powiązanie z programem Doktorat Wdrożeniowy

Niniejsza rozprawa została zrealizowana w ramach programu *Doktorat Wdrożeniowy*, projekt pt. „Elektroniczne Systemy Zabezpieczenia Technicznego jako źródło danych w procesie automatycznej obrony przed cyberatakami oraz wykrywania sprawcy przestępstwa", co implikuje tytuł rozprawy.

Niniejsza rozprawa koncentruje się na analizie i modelowaniu *Cyber-Fizycznych Systemów Ochrony (CPPS)* z wykorzystaniem metod sztucznej inteligencji, co bezpośrednio odpowiada na problemy sformułowane w temacie badawczym. Takie podejście umożliwia z jednej strony ocenę efektywności zainstalowanych systemów ochrony, poprzez wykorzystanie danych pochodzących nie tylko z samych systemów zabezpieczeń, ale również przez uwzględnienie kontekstu bezpieczeństwa i estymowanie atrybutów potencjalnego adwersarza (np. jego motywacji czy kwalifikacji). Z drugiej strony, proponowana metodologia pozwala na planowanie racjonalnego, *evidence-based* wzmocnienia środków mitygacji zagrożeń pochodzących zarówno ze sfery fizycznej, jak i cyberprzestrzeni.

W ten sposób ESZT stają się nie tylko elementem infrastruktury ochronnej, ale przede wszystkim bogatym źródłem danych, które po odpowiedniej analizie i modelowaniu stanowią fundament dla zintegrowanych strategii cyber-fizycznej obrony.

1.2. Kontekst problemu

Ochrona infrastruktury krytycznej, która obejmuje zespół powiązanych ze sobą systemów, sieci oraz aktywów fizycznych i wirtualnych, jest jednym z kluczowych elementów bezpieczeństwa narodowego. Jej awaria lub zniszczenie mogłoby bowiem poważnie zagrozić bezpieczeństwu państwa, zdrowiu publicznemu, gospodarce lub funkcjonowaniu społeczeństwa [1]. Szczególne znaczenie zyskuje ona w kontekście rosnących zagrożeń hybrydowych, które łączą w sobie ataki cybernetyczne, dezinformację oraz działania kinetyczne [2].

Aby sprostać tym wyzwaniom, Cyber-Fizyczne Systemy Ochrony (CPPS — ang. Cyber-Physical Protection Systems) integrują różnorodne środki. Są to zarówno rozwiązania techniczne, informatyczne i architektoniczne, jak również narzędzia organizacyjne, prawne oraz usługi ochrony osobowej. Ich celem jest zabezpieczenie kluczowych zasobów przed działaniami, które mogłyby destabilizować funkcjonowanie organizacji lub całych sektorów. Dotyczą one zarówno zagrożeń zamierzonych, takich jak cyberataki czy

sabotaż, jak i przypadkowych, np. awarii technicznych [3]. Rola tych systemów jest szczególnie istotna, ponieważ infrastruktura krytyczna ma interdyscyplinarną naturę. Łączy ona systemy IT i IoT z fizycznymi komponentami, takimi jak sieci energetyczne czy transportowe. Dzięki temu staje się ona priorytetowym celem dla różnych aktorów, w tym państwowych, przestępczych i terrorystycznych [4].

Ewolucja zagrożeń w ostatnich dekadach charakteryzuje się przejściem od ataków czysto fizycznych do złożonych ataków cyber-fizycznych. Historycznie, zagrożenia dla systemów bezpieczeństwa koncentrowały się wokół prób fizycznego pokonania barier — przykładowo, w sektorze bankowym dominowały ataki na bankomaty z wykorzystaniem kluczy, wiercenia czy kradzieży całych kas [5]. Współcześnie obserwujemy zasadniczą zmianę charakteru zagrożeń: przeciwnicy coraz częściej wykorzystują kombinację ataków cybernetycznych i fizycznych, używając na przykład środków wybuchowych sterowanych zdalnie czy zaawansowane ataki na systemy informatyczne kontrolujące elementy ochrony fizycznej. W sektorze bankowym transformacja ta jest szczególnie widoczna — ze względu na zmianę wiedzy i świadomości ryzyka u przeciwnika, w obszarze fizycznym dominują obecnie ataki z użyciem środków wybuchowych (przeważnie gazu ziemnego) oraz ataki cyber-fizyczne. Ta ewolucja znacząco zwiększyła ryzyko i koszty zabezpieczeń, które należy dostosować do zmiany trendu w materializacji ryzyk [6].

Paradoksalnie, mimo że nowoczesne systemy ochrony w coraz większym stopniu opierają się na technologiach informatycznych i komunikacyjnych (ICT), metodologie ich oceny pozostają zakorzenione w podejściach sprzed ery cyfrowej. Budżety przeznaczone na zabezpieczenia stanowią istotne pozycje kosztowe, dlatego stosuje się metodyki i narzędzia oceny skuteczności PPS (ang. Physical Protection Systems), które w optymalnym przypadku dostarczają informacji o realnej efektywności systemów [7]. W przeciwnym wypadku zarządzanie jest nieefektywne i może prowadzić do niegospodarności, a nawet do sytuacji, w której PPS mogą nie zrealizować celów przed nimi stawianych. Brak realizacji zadań PPS może prowadzić do poważnych konsekwencji.

1.3. Główne wyzwania i luki

Współczesne systemy ochrony fizycznej funkcjonują w coraz bardziej złożonym środowisku zagrożeń, które wykracza daleko poza tradycyjne scenariusze wtargnięć fizycznych. Digitalizacja infrastruktury krytycznej, powszechne wykorzystanie systemów IoT oraz integracja sieci teleinformatycznych z systemami bezpieczeństwa fizycznego stwarzają nowe wyzwania, które nie były uwzględniane w klasycznych metodologiach oceny ryzyka. Niniejszy rozdział identyfikuje kluczowe luki w istniejących podejściach metodologicznych, ze szczególnym uwzględnieniem:

- braku reprezentacji cyberzagrożeń w tradycyjnych modelach oceny skuteczności,
- ograniczeń w modelowaniu współczesnych scenariuszy hybrydowych,
- nieadekwatności klasycznych miar prawdopodobieństwa dla zagrożeń cyber-fizycznych,
- potrzeby nowych metodologii integrujących domenę fizyczną i cybernetyczną.

Analiza ta stanowi podstawę dla propozycji nowej metodologii przedstawionej w kolejnych rozdziałach pracy.

Brak uwzględnienia cyberzagrożeń w metodologiach oceny

Od wprowadzenia metodologii EASI (Estimated Adversary Sequence Interruption) przez Bennetta w 1977 roku [8], model ten stał się punktem wyjścia dla licznych modyfikacji i rozszerzeń w dziedzinie oceny systemów ochrony fizycznej. Metodologia EASI opiera się na fundamentalnym założeniu, że skuteczność systemu ochrony może być wyrażona poprzez prawdopodobieństwo przerwania sekwencji działań przeciwnika przed osiągnięciem celu. Klasyczna formuła EASI definiuje prawdopodobieństwo przerwania ataku $P(I)$ jako:

$$P(I) = (P_{D1}) \cdot (P_{C1}) \cdot (P(R|A_1)) + \sum_{i=2}^{n} \left[(P(R|A_i)) \cdot (P_{Ci}) \cdot (P_{Di}) \cdot \prod_{j=1}^{i-1} (1 - (P_{Dj})) \right]$$

gdzie: P_{Di} to prawdopodobieństwo detekcji podczas i-tej akcji przeciwnika, P_{Ci} to prawdopodobieństwo skutecznej komunikacji z siłami reagowania, a $P(R|A_i)$ to warunkowe prawdopodobieństwo skutecznej odpowiedzi sił reagowania. Mimo rozwoju technik ataku i postępu technologicznego, podstawowe założenia EASI nie zostały dostosowane do zagrożeń wynikających z cyfryzacji infrastruktury. Współczesne systemy PPS często wykorzystują komponenty informatyczne i sieciowe, co zwiększa ich podatność na cyberataki. Tradycyjne podejścia, w tym oryginalna wersja EASI, nie uwzględniają zagrożeń związanych z kompromitacją systemów teleinformatycznych, co może prowadzić do zbyt optymistycznej oceny skuteczności ochrony [3].

Znaczenie komponentu cybernetycznego w systemach fizycznej ochrony podkreślają także inni autorzy. Clem i Atkins analizowali przypadki ataków hybrydowych, łączących elementy fizyczne i cyfrowe, wskazując na potrzebę ich wspólnego modelowania [9]. Porter zwrócił uwagę, że odporność sieci sterujących systemami ochrony ma bezpośredni wpływ na skuteczność całego PPS [10]. Z kolei Depoy i współautorzy zaproponowali kompleksowe ramy oceny ryzyka, które łączą analizę zagrożeń fizycznych i cybernetycznych w spójnym modelu, wprowadzając wzór:

$$PE_{cp} = 1 - (1 - PE_c) \cdot (1 - PE_{pz})$$

gdzie PE_{cp} to ogólna efektywność PPS przeciwko atakom cyber-fizycznym, PE_c to prawdopodobieństwo sukcesu ataku cybernetycznego, a PE_{pz} to prawdopodobieństwo sukcesu ataku fizycznego [11]. Snell i Rivers [12] przedstawili podobne podejście dla ataków mieszanych (blended attacks) w kontekście bezpieczeństwa nuklearnego:

$$PE_{Total} = 1 - (1 - PDS) \cdot (1 - PEO)$$

gdzie PDS to prawdopodobieństwo detekcji, a PEO to szacowana ocena sukcesu dowolnego ataku fizycznego i cybernetycznego. Zastosowanie rozszerzonego podejścia, uwzględniającego zagrożenia cybernetyczne, prowadzi do istotnych różnic w ocenie skuteczności systemów ochrony. W badaniach własnych autor wykazał, że klasyczna metodologia EASI wskazywała na 89,88% prawdopodobieństwo skutecznego zakłócenia ścieżki ataku przez system ochrony. Po uwzględnieniu potencjalnych cyberataków, wartość ta została skorygowana do 1,63%, co wskazuje na potrzebę uzupełnienia klasycznych metod o komponent cybernetyczny [3].

Guo et al. [13] wykorzystali technikę Cyfrowych Bliźniaków (DT - ang. Digital Twin) do analizy ryzyka cybernetycznego PPS w elektrowniach jądrowych, tworząc dedykowaną platformę testową. Podejście to reprezentuje nowy trend w kierunku wykorzystania zaawansowanych technologii cyfrowych do modelowania i analizy systemów bezpieczeństwa. Bowen et al. [14] zaproponowali zintegrowaną platformę do analizy i projektowania PPS, umożliwiającą kompleksową ocenę efektywności systemu w warunkach operacyjnych. Andiwijayakusuma et al. [15] rozszerzyli metodę EASI o analizę wielościeżkową z uwzględnieniem zmienności parametrów przy użyciu symulacji Monte Carlo.

Problem degradacji systemów w czasie

Degradacja systemów bezpieczeństwa, w szczególności komponentów elektronicznych (ESS), jest procesem nieuniknionym, związanym z ich długoterminową eksploatacją, wpływem środowiska, a także zmianą charakterystyki zagrożeń. W kontekście CPPS degradację definiuje się jako stopniowy spadek skuteczności ochrony wynikający z czynników takich jak starzenie się komponentów, wzrost wiedzy i umiejętności przeciwnika, rozwój narzędzi ataku oraz pojawianie się nowych podatności [6]. W dotychczasowych modelach oceny skuteczności PPS, aspekt degradacji często pozostaje pomijany lub traktowany w sposób uproszczony, np. przez przyjęcie arbitralnego okresu przydatności użytkowej. Tymczasem rzeczywista skuteczność systemu ochronnego zmienia się dynamicznie w czasie, a ignorowanie tego faktu może prowadzić do zaniżonej percepcji ryzyka oraz nietrafnych decyzji inwestycyjnych. W odpowiedzi na tę lukę opracowaliśmy probabilistyczny model degradacji CPPS, który kwantyfikuje wpływ pięciu kluczowych czynników wpływających na skuteczność CPPS w czasie. Dobór rozkładów probabilistycznych dla tych czynników został oparty na charakterystyce procesów, które reprezentują, oraz na przesłankach literaturowych:

- **AG (starzenie komponentów)** – *rozkład normalny*. Starzenie jest efektem sumy wielu drobnych, w dużej mierze niezależnych mikrozjawisk (zużycie materiałów, dryf parametrów), co zgodnie z centralnym twierdzeniem granicznym uzasadnia wybór rozkładu normalnego. Podejście to znajduje potwierdzenie w literaturze dotyczącej degradacji elementów elektronicznych [16, 17].

- **KG (wzrost wiedzy przeciwnika)** – *rozkład normalny*. Akumulacja wiedzy i umiejętności następuje przyrostowo poprzez sumę wielu niezależnych aktów pozyskiwania informacji, z naturalną zmiennością w tempie i jakości gromadzonych danych, co uzasadnia zastosowanie modelowania normalnego.

- **TI (postęp technologiczny narzędzi ataku)** – *model logarytmiczny*. W początkowych fazach rozwoju narzędzi obserwuje się szybkie przyrosty efektywności, które następnie maleją zgodnie z prawem malejącej użyteczności krańcowej (*diminishing returns*). Funkcja logarytmiczna dobrze oddaje tę dynamikę technologiczną [13].

- **VI (pojawianie się nowych krytycznych podatności)** – *proces Bernoulliego*. Zdarzenia rzadkie, o charakterze skokowym i trudnej do przewidzenia częstości w krótkich horyzontach czasowych (np. podatności zero-day). W modelu ujmowane jako wystąpienie bądź brak zdarzenia w danym okresie [15, 11].

– **IE (efekt kwartalnych usprawnień)** – *rozkład normalny*. Łączny efekt wielu równoczesnych czynności serwisowych (testy, wymiany, strojenie parametrów) z naturalną zmiennością zakresu i jakości wykonywanych prac uzasadnia modelowanie sumarycznego efektu rozkładem normalnym.

Przyjęty paradygmat modelowania opiera się na zasadzie dopasowania rozkładu do natury procesu: procesy będące sumą wielu drobnych, niezależnych oddziaływań (AG, KG, IE) opisywane są rozkładem normalnym; zdarzenia rzadkie i skokowe (VI) – procesem Bernoulliego; natomiast procesy, w których kolejne innowacje przynoszą coraz mniejsze przyrosty efektywności (tzw. *diminishing returns*), jak w przypadku rozwoju narzędzi używanych do ataku (TI), najlepiej odwzorowuje funkcja logarytmiczna. Poniżej przedstawione wartości wspomnianych czynników zostały dobrane tak, aby odzwierciedlały aktualne podejście praktyków, jednak stanowią punkt wyjścia dla dalszych badań i ich precyzyjniejszej kalibracji:

1. **Starzenie się komponentów (AG)** - modelowane rozkładem normalnym (np. $\mu = 0{,}5$, $\sigma = 0{,}15$), co odzwierciedla średni spadek sprawności o około 50 punktów procentowych w dekadzie (rozrzut $\pm$15 pp) [16, 17].

2. **Wzrost wiedzy przeciwnika (KG)** - rozkład normalny (np. $\mu = 0{,}1$, $\sigma = 0{,}03$) odpowiadający około 10% rocznemu przyrostowi skuteczności ataków dzięki lepszemu planowaniu i dostępowi do wiedzy [11].

3. **Postęp technologiczny narzędzi atakujących (TI)** - model logarytmiczny (np. $\mu = 0{,}7, \sigma = 0{,}21$) odzwierciedlający przyspieszenie innowacji, gdzie kolejne usprawnienia przynoszą coraz mniejszy przyrost wydajności [13].

4. **Pojawianie się nowych podatności (VI)** - proces Bernoulliego (np. $p = 0{,}2$, jednorazowy spadek = 10 pp) reprezentujący pojawienie się nowych, krytycznych luk z historyczną częstością około 1 na 5 miesięcy [15].

5. **Efekty kwartalnych usprawnień (IE)** - rozkład normalny (np. $\mu = 0{,}5$, $\sigma = 0{,}15$) modelujący prewencyjne działania konserwacyjne (testy, wymiana części) co 3 miesiące, redukujące skumulowany spadek efektywności o średnio 50 pp z odchyleniem $\pm$15 pp.

Model zakłada obliczanie skuteczności systemu w danym miesiącu $SE_i = \max(100 - X_i, 0)$, gdzie X_i to skumulowany wpływ degradacyjny [6]. Wybór rozkładu normalnego dla większości parametrów wynika z centralnego twierdzenia granicznego - wiele drobnych czynników wpływających na procesy starzenia czy konserwacji sumuje się do rozkładu normalnego. Model logarytmiczny dla postępu technologicznego odzwierciedla charakterystykę „malejących przychodów” typową dla rozwoju technologii. Proces Bernoulliego dla podatności oddaje rzadkie, ale skokowe zdarzenia krytyczne.

Symulacje Monte Carlo przeprowadzone dla 10,000 przebiegów wykazały, że efektywność typowego systemu może spaść z poziomu 98% do około 25% w horyzoncie 10 lat. Co istotne, średnia linia trendu przekracza poziom 50% około 75.–80. miesiąca eksploatacji, co oznacza, że dziesięcioletnie cykle wymiany sprzętu mogą być zbyt długie, aby zapewnić adekwatną ochronę [6].

Studium przypadku - wpływ degradacji na ocenę EASI

Praktyczne konsekwencje ignorowania degradacji można zilustrować na przykładzie typowej ścieżki ataku obejmującej cztery warstwy ochronne:

- **Ochrona obwodowa** – sensometryczny przewód światłowodowy monitorujący linię ogrodzenia (2 lata eksploatacji)
- **Nadzór wizyjny** – kamery CCTV nadzorujące obszar chroniony (7 lat eksploatacji)
- **Kontrola dostępu** – czytniki transponderów RFID oraz czujniki otwarcia (kontaktrony) (5 lat eksploatacji)
- **System Sygnalizacji Włamania** – pasywne czujki podczerwieni (PIR) wykrywające ruch wewnątrz obiektu (10 lat eksploatacji)

Klasyczna ocena EASI wskazywała prawdopodobieństwo przerwania ataku na poziomie 88,64%. Po uwzględnieniu współczynników degradacji SE_i dla każdego elementu, skuteczność spadła do 76,19% - różnica 12,45 punktów procentowych może zadecydować o powodzeniu lub niepowodzeniu ataku [6]. Włączenie wskaźnika SE_i do rozszerzonego modelu EASI skutkowało spadkiem estymowanego prawdopodobieństwa skutecznej detekcji ścieżki ataku o 15–25 punktów procentowych już po pięciu latach działania systemu, co ilustrują rysunki 2.2 i 2.3.

Analiza poszczególnych składników X_i (linia pomarańczowa, zielona, fioletowa, magenta i brązowa linia) pokazuje wyraźne, nierówne udziały (patrz rys. 1.3):

- Starzenie się (AG_j, pomarańczowa) rośnie liniowo od 0 do około 60–80 punktów procentowych w dekadzie i stanowi największy udział w skumulowanym spadku.
- Wzrost wiedzy przeciwnika (KG_j, zielona) narasta stopniowo od 0 do około 10–15 pp w ciągu dziesięciu lat.
- Udoskonalanie narzędzi (TI_j, fioletowa) przyspiesza w drugiej połowie okresu i sięga do ok. 8–10 pp.
- Podatności (VI_j, magenta) dają sporadyczne skoki spadku do 5–10 pp, ale są rzadkie (średnio jeden na pięć miesięcy).
- Efekty konserwacji (IE_j, brązowa) pojawiają się kwartalnie i mają niewielką wartość – rzędu 1–2 pp regeneracji na cykl, co jest niewystarczające, by skompensować skumulowane straty.

Taki wynik jednoznacznie wskazuje na potrzebę uzupełnienia klasycznych metod oceny skuteczności CPPS o dynamiczne podejścia uwzględniające zmienność parametrów ochronnych w czasie.

Luka w modelowaniu degradacji ESS

Przegląd badań w omawianym zakresie wskazał dwie główne grupy prac dotyczące degradacji systemów. Pierwsza grupa koncentruje się na analizie starzenia się elementów elektronicznych:

Response Relaybility	Mean time of response (sec.)	Force Time standard dev. (sec.)
0,95	250	90

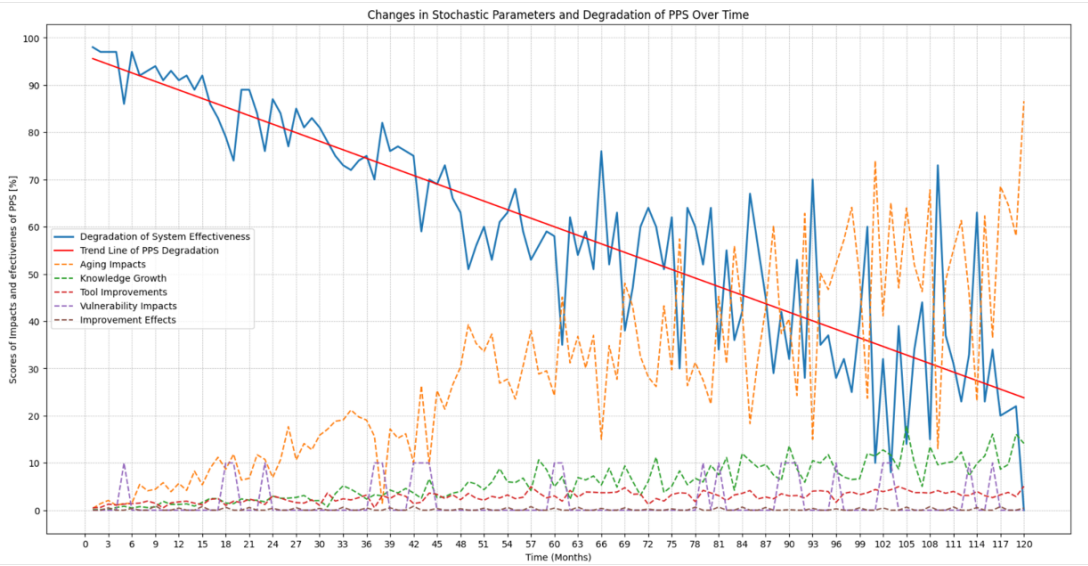
Task #:	Adversary task description	Pd (probability of detection)	Detection Location (B/M/E)	mean delay	sdev	Normal values	cum PI
1	cut fence	90%	B	90	3	0,950385092	0,8125793
2	run	90%	B	12	3,6	0,784149763	0,0670448
3	open a door	90%	B	90	27	0,749086947	0,0064047
4	run	90%	B	10	3	0,421259148	0,0003602
5	open a door	90%	B	100	27	0,382797242	3,273E-05
6	sabotage of the asset	90%	B	120	36	0,089938638	7,69E-07
7		0%	0	0	0	0,002736602	0
8		0%	0	0	0	0,002736602	0
9		0%	0	0	0	0,002736602	0
10		0%	0	0	0	0,002736602	0
11		0%	0	0	0	0,002736602	0
12		0%	0	0	0	0,002736602	0
						Probability of interuption (Physical attack)	0,8864

Rysunek 1.1: Ocena skuteczności PPS metodą EASI dla czysto fizycznego wektora ataku bez uwzględnienia degradacji komponentów. Tabela przedstawia nominalne prawdopodobieństwa detekcji (Pd = 90%) dla wszystkich elementów systemu, podając w rezultacie prawdopodobieństwo przerwania ataku na poziomie 88,64%. Źródło: opracowanie własne na podstawie [6].

Response Relaybility	Mean time of response (sec.)	Force Time standard dev. (sec.)
0,95	250	90

Task #:	Adversary task description	Detection Element/s	Age of system	Mean of estimated efectivenes from degradation model	Probability of detection on the installation day	Pd (probability of detection on the day of evaluation)	Detection Location (B/M/E)	mean delay	sdev	Normal values	cum PI
1	cut fence	Fiber-optic sensor	3	70%	90%	63%	B	90	3	0,950385092	0,5688055
2	run	CCTV camera	7	30%	90%	27%	B	12	3,6	0,784149763	0,0744197
3	open a door	Door contact	5	50%	90%	45%	B	90	27	0,749086947	0,0864954
4	run	PIR Detector	8	20%	90%	18%	B	10	3	0,421259148	0,0107012
5	open a door	Door contact	5	50%	90%	45%	B	100	27	0,382797242	0,0199345
6	sabotage of the asset	CCTV camera	7	30%	90%	27%	B	120	36	0,089938638	0,0015456
7						0%	0	0	0	0,002736602	0
8						0%	0	0	0	0,002736602	0
9						0%	0	0	0	0,002736602	0
10						0%	0	0	0	0,002736602	0
11						0%	0	0	0	0,002736602	0
12						0%	0	0	0	0,002736602	0
										Probability of interuption (Physical attack)	0,7619

Rysunek 1.2: Ocena skuteczności PPS metodą EASI z uwzględnieniem degradacji komponentów w czasie eksploatacji. Kolumna “Age of system” określa wiek poszczególnych elementów, a kolumna “Mean of estimated effectiveness from degradation model” pokazuje efektywność po uwzględnieniu procesu starzenia. Końcowe prawdopodobieństwo przerwania ataku wynosi 76,19%. Źródło: opracowanie własne na podstawie [6].



Rysunek 1.3: Wykres wpływu uwzględnionych w modelu degradacji czynników na efektywność systemów bezpieczeństwa. Źródło: opracowanie własne na podstawie [6].

Narale et al. [16] przeprowadzili dogłębną analizę struktury i mechanizmów degradacji kondensatorów elektrolitycznych, wykorzystując charakterystyki materiałowe. Yu et al. [17] zbadali symulacyjnie wpływ hot-carrier stress na wydajność konwerterów DC–DC, identyfikując główne czynniki prowadzące do degradacji. Alam i Khan [18] opracowali model niezawodności i wydajności konwertera Boost, określając, w jaki sposób zużycie komponentów wpływa na spadek parametrów. Huang et al. [19] zaproponowali model degradacji pasywnych elementów w długoterminowej analizie emisji elektromagnetycznej.

Druga grupa badań koncentruje się na metodach prognozowania degradacji różnych typów systemów:

– **Modelowanie stochastyczne i statystyczne**: Du et al. [20] zaproponowali metodę estymacji pozostałego czasu życia systemów o dużych fluktuacjach degradacji. Fang et al. [21] zastosowali kopułę (Copula) do analizy niezawodności systemów ze współzależnymi awariami.

– **Reinforcement Learning**: Dong et al. [22] przedstawili przegląd podstaw uczenia przez wzmacnianie oraz ich zastosowań w optymalizacji decyzji dla systemów degradacyjnych. Hua [23] opracował podejście wykorzystujące uczenie przez wzmacnianie do optymalizacji wydajności systemów ze sprzężeniem zwrotnym.

– **C**yfrowy Bliźniak: Peng et al. [24] wprowadzili metodę cyfrowego bliźniaka do estymacji wskaźników zdrowia konwerterów DC–DC. Chen et al. [25] zaproponowali cyfrowego bliźniaka z optymalizacją Bayesowską do identyfikacji parametrów degradacji przetwornic.

Żaden z analizowanych modeli nie odnosi się bezpośrednio do degradacji ESS w kontekście CPPS. Istniejące podejścia koncentrują się na konkretnych komponentach elektronicznych lub ogólnych systemach technicznych, ale nie uwzględniają specyfiki systemów bezpieczeństwa i ich złożonych interakcji z środowiskiem operacyjnym.

Problem standaryzacji symboli i interpretacji kontekstu

Kolejnym zidentyfikowanym elementem wymagającym opracowania jest detekcja elementów ESZT na planach uwzględniających ich fizyczną lokalizację. Element ten jest niezbędny dla dalszych prac implementacyjnych, które opierać się będą na analizie grafów. Automatyczne rozpoznawanie i lokalizacja symboli technicznych na planach architektonicznych stanowi ważny element dalszej analizy systemów bezpieczeństwa. Tradycyjne podejścia oparte na dopasowaniu wzorców czy algorytmach morfologicznych napotykają trudności związane z różnorodnością stylów rysowania, skal przedstawienia oraz jakości skanowanych dokumentów. W związku z tym uwagę skierowano na nowoczesne architektury głębokich sieci neuronowych. Do tego zadania wybraliśmy architekturę YOLO (You Only Look Once) [26], która traktuje detekcję obiektów jako problem jednoetapowej regresji współrzędnych ramek ograniczających i prawdopodobieństw klas. Zalety tego podejścia w kontekście analizy planów technicznych obejmują: wysoką szybkość przetwarzania, pozwalającą na analizę dużych arkuszy planów, możliwość jednoczesnej detekcji wielu obiektów różnych klas, oraz dobrą adaptację do obiektów o małych rozmiarach, charakterystycznych dla symboli na planach architektonicznych.

Konwencjonalne modele YOLO, choć sprawnie lokalizują i klasyfikują pojedyncze symbole, stanowią punkt wyjścia dla pełnej analizy planów bezpieczeństwa. Dalszym wyzwaniem pozostaje zrozumie-

nie strukturalnych relacji pomiędzy wykrytymi elementami oraz ich funkcjonalnej semantyki, czyli tego, co dany symbol oznacza w kontekście otaczających go obiektów oraz szerszej logiki systemu bezpieczeństwa [27]. Problematyka ta objawia się na trzech płaszczyznach:

- **Niejednoznaczność wizualna symboli** — ten sam znak graficzny może oznaczać różne urządzenia w zależności od konwencji projektowej, standardu architektonicznego lub domeny zastosowania. Przykładowo, prostokąt z ukośną kreską może oznaczać zarówno drzwi przesuwne, jak i zabezpieczenie przeciwpożarowe, w zależności od projektu.
- **Brak kontekstualizacji w procesach CV** — tradycyjne podejścia separują detekcję od interpretacji. Oznacza to, że nawet precyzyjna detekcja symboli nie gwarantuje poprawnego zrozumienia ich roli w systemie.
- **Kruchość reguł heurystycznych** — interpretacja znaczenia symboli często opiera się na ręcznie tworzonych regułach, które są trudne do uogólnienia i podatne na błędy w przypadkach niestandardowych lub złożonych układów.

Ewolucja metod detekcji obiektów

Rozwój systemów komputerowej analizy obrazów zapewnia istotny kontekst dla zrozumienia wkładu COG. Klasyczne podejścia opierały się na ręcznie tworzonych cechach, z SIFT [28] ustanawiającym zasady ekstrakcji cech niezależnych od skali. Rewolucja deep learning rozpoczęła się od AlexNet [29], który wykazał moc konwolucyjnych sieci neuronowych, a następnie ulepszenia architektoniczne w VGG [30] i ResNet [31].

Nowoczesna detekcja obiektów pojawiła się wraz z R-CNN [32], które wprowadziło detekcję opartą na regionach poprzez "selective search"i ekstrakcję cech CNN. YOLO [26] zrewolucjonizowało dziedzinę, formułując detekcję jako pojedynczy problem regresji.

Najnowsze osiągnięcia w kontekstowej detekcji obiektów, w tym YOLO-World [33] z integracją vision-language oraz Florence-2 [34] z zunifikowanymi możliwościami multi-task, demonstrują znaczący postęp w rozumieniu kontekstowym. YOLO-World pozwala na detekcję otwartych słowników poprzez integrację wizyjnych i językowych embeddingów bezpośrednio w procesie detekcji. Florence-2 posuwa to dalej, unifikując różne zadania vision pod jedną reprezentacją.

Analiza Układu Dokumentów

Kluczowym wyzwaniem jest również brak standaryzacji w reprezentacji elementów infrastruktury. Jak wykazały badania nad systemami analizy układu dokumentów (ang. DLA - Document Layout Analysis), najnowsze osiągnięcia takie jak LayoutLM [35] wprowadziły pre-training tekstu i layout dla rozumienia obrazów dokumentów. LayoutLMv3 [36] osiągnął state-of-the-art performance poprzez zunifikowane maskowanie tekstu i obrazu, demonstrując możliwości integracji wizyjnych i językowych embeddingów dla lepszego rozumienia kontekstu. Jednak w dziedzinie diagramów bezpieczeństwa fizycznego brakuje dedykowanych rozwiązań łączących detekcję z interpretacją funkcjonalną [27].

W odpowiedzi na te wyzwania opracowaliśmy podejście COG, które integruje klasyczną detekcję obiektów z analizą relacji topologicznych oraz uwzględnia kontekst funkcjonalny otoczenia. Zamiast trak-

tować symbole jako niezależne jednostki, COG grupuje je zgodnie z ich przestrzennym i semantycznym sąsiedztwem, umożliwiając bardziej realistyczne odwzorowanie struktury systemu bezpieczeństwa. Takie podejście stanowi fundament dla dalszych etapów modelowania grafów ataku i oceny ryzyka w CPPS [27].

Zastosowania AI/ML w bezpieczeństwie

Zastosowanie sztucznej inteligencji w systemach bezpieczeństwa rozwija się w kierunku inteligentnych środowisk detekcji. Li et al. [37] przedstawili przegląd tzw. smart sensing w budownictwie, wskazując na obecny stan i przyszłe perspektywy. Wang et al. [38] analizowali tzw. 'emerging IoT technologies' dla 'smart' środowisk.

Kompleksowe przeglądy [39, 40] ustanawiają znaczenie automatyzacji oceny bezpieczeństwa w automatyzacji budynków i smart cities. Yohanandhan et al. [40] przedstawili kompleksowy przegląd cyberfizycznych systemów i smart cities w Indiach, identyfikując możliwości, problemy i wyzwania. Autorzy podkreślili znaczenie automatyzacji oceny bezpieczeństwa w kontekście inteligentnych środowisk detekcji.

Najnowsze prace w MDPI Sensors demonstrują praktyczne zastosowania obejmujące smart sensing w konstrukcji budynków [37] i środowiska IoT-based smart [38].

1.4. Konsekwencje obecnego stanu

Niedoskonałości obecnych metod oceny CPPS mają istotne implikacje praktyczne dla zarządzania bezpieczeństwem infrastruktury krytycznej. Luki związane z nieuwzględnianiem degradacji, cyberzagrożeń oraz kontekstu operacyjnego skutkują poważnymi ryzykami dla organizacji, zarówno w wymiarze technicznym, jak i ekonomicznym.

Po pierwsze, systematyczne przeszacowanie skuteczności systemów ochrony prowadzi do błędnych decyzji inwestycyjnych. Jeśli model oceny opiera się na nierealistycznie wysokim poziomie ochrony (np. nie uwzględnia spadku efektywności elementów czy podatności na ataki cybernetyczne), to inwestycje w modernizację mogą być odkładane, mimo że system w rzeczywistości nie spełnia już wymagań bezpieczeństwa [5]. Na podstawie porównania obu ocen widać, że uwzględnienie degradacji obniża znaczniki skuteczności CPPS o dodatkowe 15–25 punktów procentowych już w 5. roku użytkowania. W klasycznym podejściu (bez degradacji) CPPS wygląda na relatywnie znacznie bardziej odporne, co może prowadzić do błędnych decyzji, w tym o zbyt rzadkiej wymianie sprzętu.

Po drugie, ignorowanie zjawiska degradacji prowadzi do utraty kontroli nad realną skutecznością systemów w czasie. Brak danych o aktualnym stanie technicznym komponentów uniemożliwia trafne planowanie przeglądów, wymian czy uzupełnień infrastruktury ochronnej. W efekcie dochodzi do utrzymywania sprzętu, który mimo formalnej sprawności nie zapewnia już efektywnej ochrony. Choć pojedyncze realizacje modelu (niebieska linia) czasem schodzą poniżej 50% już po 60 miesiącach, średnia tendencja (linia trendu) wskazuje, że poziom 50% zostaje przekroczony dopiero około 75–80 miesiąca

(6,5–6,7 roku). Oznacza to, że w rozważanym scenariuszu typowy cykl wymiany CPPS krótszy niż 7 lat byłby uzasadniony, jeśli chcemy utrzymać efektywność powyżej połowy początkowej.

Po trzecie, nieuwzględnianie kontekstu operacyjnego i sytuacyjnego może skutkować zastosowaniem rozwiązań ochronnych niedostosowanych do charakterystyki chronionego obiektu. Na przykład system zaprojektowany z myślą o ochronie magazynu może nie być skuteczny w przypadku obiektu wysokiego ryzyka (np. infrastruktury energetycznej lub placówki dyplomatycznej), jeśli nie uwzględniono uwarunkowań takich jak potencjalna aktywność przeciwnika, lokalizacja geograficzna czy otoczenie społeczne.

Ponadto, brak automatyzacji i standaryzacji w zakresie detekcji oraz analizy danych skutkuje dużym uzależnieniem od indywidualnych ocen ekspertów. Procesy te bywają czasochłonne i kosztowne, a także trudne do powtórzenia czy audytowania. W dłuższej perspektywie może to prowadzić do niespójnych decyzji oraz trudności w porównywaniu efektywności różnych rozwiązań ochronnych. Analiza poszczególnych składników degradacji pokazuje wyraźne, nierówne udziały - starzenie się stanowi największy udział w skumulowanym spadku, podczas gdy kwartalne usprawnienia dają jedynie niewielką wartość regeneracji rzędu 1–2 pp na cykl, co jest niewystarczające, by skompensować skumulowane straty.

Łącznie, opisane niedoskonałości zwiększają ryzyko kompromitacji systemu ochrony, co może prowadzić do incydentów naruszających bezpieczeństwo ludzi, infrastruktury oraz danych. W kontekście rosnącej presji regulacyjnej oraz społecznego zainteresowania odpornością na zagrożenia hybrydowe, konieczne jest wypracowanie bardziej wiarygodnych, dynamicznych i kontekstowych metod oceny skuteczności CPPS.

1.5. Identyfikacja luk badawczych

Główne zidentyfikowane luki

Na podstawie przeprowadzonej analizy stanu wiedzy zidentyfikowaliśmy następujące kluczowe luki badawcze:

1. **Brak integracji cyber-fizycznej**: Mimo rozpoznania problemu przez badaczy takich jak Clem i Atkins [9], Porter [10] czy Depoy et al. [11], brakuje kompleksowych, praktycznych rozwiązań integrujących ocenę cyberzagrożeń z metodologiami oceny bezpieczeństwa fizycznego. Istniejące propozycje są często teoretyczne i nie uwzględniają złożoności rzeczywistych systemów.

2. **Nieuwzględnienie degradacji w czasie**: Istniejące metodologie zakładają stałą efektywność systemów, ignorując rzeczywiste procesy degradacji wpływające na skuteczność ochrony. Mimo że w dziedzinie inżynierii elektronicznej istnieją zaawansowane modele degradacji komponentów [16, 17], żaden z nich nie odnosi się bezpośrednio do degradacji ESS w kontekście CPPS.

3. **Ograniczenia w interpretacji kontekstu**: Tradycyjne systemy komputerowej analizy obrazów nie radzą sobie z dynamiczną interpretacją symboli w diagramach technicznych, gdzie znaczenie zależy od kontekstu lokalnego. Mimo postępu w dziedzinie DLA [35, 36], brakuje dedykowanych rozwiązań dla diagramów bezpieczeństwa.

4. **Brak automatyzacji kompleksowej oceny**: Obecne podejścia wymagają znacznej interwencji ekspertów i nie zapewniają zautomatyzowanych, skalowalnych rozwiązań. Mimo rozwoju inteligentnych systemów detekcji [37, 38], ich integracja z metodologiami oceny bezpieczeństwa pozostaje ograniczona.

5. **Fragmentaryczność podejść**: Istniejące rozwiązania adresują pojedyncze aspekty problemu, brakuje holistycznego, zintegrowanego procesu łączącego wszystkie wymiary oceny CPPS.

Uzasadnienie potrzeby nowych badań

Zidentyfikowane luki wskazują na potrzebę opracowania nowych, zintegrowanych podejść do oceny CPPS. Szczególnie istotne jest:

– Stworzenie metodologii łączącej aspekty cyber-fizyczne w jednej spójnej metodologii oceny

– Opracowanie modeli degradacji specyficznych dla systemów bezpieczeństwa elektronicznego

– Wykorzystanie zaawansowanych technik AI/ML do automatyzacji procesów interpretacji i oceny

– Integracja różnych źródeł danych i kontekstu w dynamicznym systemie oceny

Integracja wskaźnika degradacji SE_i z metodą EASI pozwala na uzyskanie bardziej realistycznych i konserwatywnych estymat prawdopodobieństwa przerwania ataku, obniżając je o dodatkowe 15–25 punktów procentowych już po pięciu latach eksploatacji. Taka modyfikacja wydaje się być istotną daną z punktu widzenia planowania budżetowego i zabezpieczenia ciągłości działań ochronnych.

1.6. Cel i zakres pracy

W odpowiedzi na zidentyfikowane wcześniej ograniczenia klasycznych metod oceny systemów ochrony, celem niniejszej rozprawy jest opracowanie zintegrowanego, opartego na danych modelu oceny skuteczności cyber-fizycznych systemów bezpieczeństwa (CPPS). Proponowany model uwzględnia dynamikę degradacji komponentów technicznych, rosnącą rolę zagrożeń cybernetycznych oraz wpływ kontekstu operacyjnego, w którym funkcjonuje dany system.

Założeniem pracy jest, że skuteczna ocena CPPS wymaga integracji trzech kluczowych obszarów:

– degradacji komponentów i infrastruktury ochronnej w czasie [6],

– wpływu zagrożeń cybernetycznych na elementy fizyczne [41],

– specyfiki operacyjnej i kontekstualnej chronionego obiektu.

Postawiono tezę, że klasyczne metody (np. EASI) są niewystarczające w długoterminowej ocenie skuteczności CPPS, a ich uzupełnienie o dane techniczne, degradacyjne i cybernetyczne prowadzi do bardziej realistycznej i trafnej oceny ryzyka. Ponadto automatyzacja procesu detekcji elementów infrastruktury

oraz analizy ścieżek ataku z wykorzystaniem algorytmów heurystycznych umożliwia skalowalne zastosowania w praktyce [42, 43].

Zakres pracy obejmuje:

1. Opracowanie probabilistycznego modelu degradacji CPPS oraz jego parametryzację na podstawie danych eksperckich i symulacyjnych.
2. Implementację procesu COG umożliwiającego automatyczną interpretację planów technicznych i symboli zabezpieczeń.
3. Zastosowanie algorytmu A* z kontekstową funkcją kosztu do oceny ścieżek ataku przeciwnika.
4. Integrację dużych modeli językowych (LLM) do analizy kontekstu operacyjnego i generowania zrozumiałych raportów ryzyka.
5. Walidację modelu na rzeczywistych przypadkach testowych, z porównaniem skuteczności podejścia klasycznego i zintegrowanego.

Model wymaga kalibracji na danych z rzeczywistych instalacji CPPS, zwłaszcza w odniesieniu do kwartalnych efektów konserwacji i częstotliwości krytycznych podatności. Dopiero wówczas będzie możliwe opracowanie precyzyjnych, ekonomicznie zoptymalizowanych harmonogramów konserwacji i wymiany komponentów. Jako naturalną kontynuację tych badań proponujemy cztery główne obszary rozwoju:

Po pierwsze, w zakresie innowacji technologicznych warto zbadać wykorzystanie algorytmów AI i AI do predykcji krytycznych momentów degradacji, a także eksperymenty z technologią blockchain dla zapewnienia niezmienności i audytowalności danych pomiarowych. Integracja czujników IoT nowej generacji, takich jak urządzenia akustyczne czy termowizyjne, pozwoli natomiast na wczesne wykrywanie oznak zużycia.

Po drugie, rekomendujemy opracowanie dynamicznych polityk i regulacji, które na bieżąco aktualizują wymagania techniczne CPPS w oparciu o wskaźniki degradacji i ryzyka. Zasadnym wydaje się również rozważenie współpracy międzynarodowej nad ujednoliceniem progów efektywności i cykli życia systemów ochrony oraz analizę wpływu lokalnych norm. Tego rodzaju podejście mogłoby wzmocnić spójność praktyk utrzymania i modernizacji.

Po trzecie, zasadne jest pogłębienie badań nad strategiami organizacyjnymi. Wdrożenie kultury proaktywnej konserwacji wymaga zarówno szkoleń technicznych dla personelu, jak i wsparcia ze strony kadry zarządzającej, a także analiz międzysektorowych, które wskażą najlepsze praktyki w branżach takich jak finanse, energetyka czy transport.

Wreszcie, kluczowe znaczenie wydają się mieć badania długoterminowe. Wieloletnie gromadzenie rzeczywistych danych eksploatacyjnych z instalacji CPPS, w tym pomiarów efektywności po każdej interwencji serwisowej, oraz połączenie ich z analizą ekonomiczną kosztów konserwacji, modernizacji i ryzyka awarii. Taka baza danych umożliwi precyzyjną kalibrację modelu w różnych warunkach środowiskowych (temperatura, wilgotność, natężenie ruchu) i pozwoli na regionalizację zaleceń dotyczących utrzymania.

1.7. Teza pracy

Teza pracy została sformułowana następująco:

Możliwe jest opracowanie zintegrowanego, probabilistycznego modelu oceny skuteczności Cyber-Fizycznych Systemów Ochrony (CPPS), który jednocześnie uwzględniał będzie: degradację komponentów technicznych w czasie, zagrożenia cybernetyczne oraz operacyjny kontekst funkcjonowania systemu.

1.8. Struktura rozprawy

W dalszej części rozprawy:

Rozdział 2 prezentuje przegląd metod oceny systemów bezpieczeństwa fizycznego oraz identyfikuje kluczowe deficyty w istniejących podejściach, ze szczególnym uwzględnieniem problematyki degradacji systemów w czasie. Rozdział 3 omawia zastosowanie algorytmów komputerowej analizy obrazów, w szczególności YOLO, do automatycznej detekcji elementów infrastruktury budynków oraz ich post-processingu w kontekście analizy bezpieczeństwa. Rozdział 4 przedstawia metodologię generowania grafów ścieżek ataku oraz heurystyczną ocenę tych ścieżek z wykorzystaniem algorytmów przeszukiwania grafów i analizy topologicznej. Rozdział 5 opisuje zastosowanie dużych modeli językowych (LLM) w syntezie danych bezpieczeństwa oraz automatycznym generowaniu raportów i rekomendacji. Rozdział 6 przedstawia kompleksową implementację proponowanego podejścia w formie funkcjonalnego narzędzia oraz przykłady jego zastosowania w rzeczywistych scenariuszach oceny systemów bezpieczeństwa. Rozdział 7 podsumowuje wnioski oraz analizuje ograniczenia i dalsze kierunki rozwoju proponowanego podejścia.

1.9. Podsumowanie

Niniejsza praca przedstawia zintegrowane, oparte na danych podejście do oceny skuteczności cyber-fizycznych systemów ochrony (CPPS), łączące trzy komplementarne perspektywy: dynamikę degradacji elementów technicznych, oddziaływanie zagrożeń cybernetycznych na składowe fizyczne oraz kontekst operacyjny i topologiczny chronionego obiektu. Rdzeń rozwiązania stanowi probabilistyczny model ujęcia pięciu klas czynników (starzenie komponentów, wzrost kompetencji przeciwnika, postęp narzędzi ataku, pojawianie się podatności oraz efekty działań utrzymaniowych), rozszerzony o komponent cyber i sprzężony z oceną ścieżek ataku. Warstwa percepcyjna wykorzystuje kontekstową analizę planów technicznych i detekcję obiektów (COG), warstwa decyzyjna używa grafów ścieżek ataku z heurystyką A* oraz modyfikację metody EASI o wskaźniki degradacji i element ryzyka pochodzący z cyberprzestrzeni, a całość dopełnia automatyzacja raportowania z użyciem modeli językowych oraz otwarcie na integrację z ekosystemem IoT/Digital Twin.

Symulacje Monte Carlo i studia przypadków pokazują, że wspólne modelowanie tych komponentów istotnie zmienia ocenę ryzyka: ujęcie czynników degradacyjnych i cybernetycznych obniża prognozowaną skuteczność detekcji o kilkanaście–kilkadziesiąt punktów procentowych, a w analizowanym scenariuszu średni poziom efektywności CPPS spada poniżej 50% po 6,5–7 latach eksploatacji. Wyniki te

uzasadniają odejście od statycznych harmonogramów na rzecz dynamicznego, kontekstowego zarządzania cyklem życia systemów oraz stanowią podstawę do wdrożenia prototypu i jego kalibracji na danych rzeczywistych.

2. Zasady oceny systemów bezpieczeństwa fizycznego i zidentyfikowane deficyty

Jak wyjaśniono w poprzednim rozdziale, współczesne systemy ochrony fizycznej (PPS) stoją przed fundamentalnymi wyzwaniami wynikającymi z cyfryzacji infrastruktury (ICT), postępującej degradacji komponentów elektronicznych oraz rosnącej złożoności zagrożeń o charakterze hybrydowym. Klasyczne metodologie oceny, choć stanowiły przez dekady solidny fundament zarządzania bezpieczeństwem, nie nadążają za tymi zmianami.

Niniejszy rozdział koncentruje się na praktycznych aspektach oceny systemów bezpieczeństwa. Przedstawia krótkie przypomnienie klasycznej metodologii EASI jako punktu wyjścia, precyzyjnie identyfikuje jej kluczowe deficyty w kontekście współczesnych zagrożeń, a następnie szczegółowo opisuje autorskie rozwiązania adresujące te luki: rozszerzenie EASI o komponent cyberodporności (CR) oraz model probabilistyczny degradacji systemów. Rozdział kończy się zapowiedzią rozwiązań automatyzujących proces oceny, które zostaną szczegółowo omówione w kolejnych częściach rozprawy.

2.1. Metodologia EASI - przypomnienie

EASI (Estimated Adversary Sequence Interruption) wprowadzona przez Bennetta w 1977 roku [8] pozostaje de facto standardem oceny skuteczności PPS. Model definiuje prawdopodobieństwo przerwania ataku $P(I)$ jako:

$$P(I) = (P_{D1}) \cdot (P_{C1}) \cdot (P(R|A_1)) + \sum_{i=2}^{n} \left[(P(R|A_i)) \cdot (P_{Ci}) \cdot (P_{Di}) \cdot \prod_{j=1}^{i-1} (1 - (P_{Dj})) \right]$$

gdzie:

- P_{Di} - prawdopodobieństwo detekcji podczas i-tej akcji przeciwnika
- P_{Ci} - prawdopodobieństwo skutecznej komunikacji z siłami reagowania
- $P(R|A_i)$ - warunkowe prawdopodobieństwo skutecznej odpowiedzi sił reagowania

EASI opiera się na trzech fundamentalnych założeniach:

1. Przeciwnik działa wyłącznie fizycznie

2. Skuteczność komponentów pozostaje stała w czasie
3. Analiza wymaga ręcznego modelowania ścieżek ataku

Te założenia, choć uzasadnione w latach 70., dziś stanowią główne ograniczenia metodologii.

2.2. Zidentyfikowane deficyty klasycznych podejść

Brak uwzględnienia zagrożeń cybernetycznych

Współczesne PPS wykorzystują rozbudowane sieci ICT do sterowania detekcją, komunikacją i reakcją. Każdy cyfrowy komponent może stać się celem cyberataku, który może:

– Wyłączyć system detekcji
– Zafałszować dane z czujników
– Zablokować komunikację
– Otworzyć elektronicznie sterowane zamki

Znaczący wpływ zagrożeń cybernetycznych na skuteczność systemów ochrony fizycznej ilustruje rysunek 2.1. Porównanie przedstawia ocenę tej samej ścieżki ataku przy użyciu klasycznej metodologii EASI oraz zmodyfikowanej wersji uwzględniającej cyber-odporność elementów systemu.

							Cyber effectiveness of communication	Description of the impact of cyber-factor
Response Time (s)	360							no impact
Response Time Std. Deviation (s)	36							no impact
Detection / Response Reliability (%)	89,00%						90%	assuming 10% of chances tu distrurb communication
Description of step in adversary sequence	Probability of Detection Pd [%]	Cybersec. Effectiveness of PPS element Pec[%]	Overall Probability of detection PDT=PdPec	Delay (s)	Delay Std. Deviation (s)	Delay impacted by cyber-attack (s)	Detection type	Description of cyber-impact (if the cyber-controlled element is impacted in its Detection Probability / its Time of delay or both?
Site Perimeter (Defeat Fence)	90,00%	10%	9%	60	6	60	B	impact od detection only
Run to the building wall	90,00%	10%	9%	30	3	30	B	impact od detection only
Defeat window	90,00%	10%	9%	90	9	10	B	impact on both (delay time reduced to 10 sec.)
Walk to inner-perimeter door	90,00%	10%	9%	50	5	50	B	impact od detection only
inner-perimeter door	90,00%	10%	9%	10	18	10	B	impact on both (time reduced to 10 sec.)
Walk to door to "target-room"	90,00%	10%	9%	40	4	40	B	impact od detection only
Defeat door to "target-room"	90,00%	10%	9%	10	5	50	B	impact od detection only
Actions on a target	90,00%	10%	9%	30	3	30	B	impact od detection only
Probability of Adversary Sequence Interruption					89,88%			
Probability of Adversary Sequence Interruption considering Cyber-attack on PPS					1,63%			

Rysunek 2.1: Porównanie oceny skuteczności PPS metodą EASI: klasycznej (89,88% prawdopodobieństwa przerwania ataku) z uwzględnieniem cyberodporności (1,63% prawdopodobieństwa przerwania ataku). Tabela pokazuje wpływ redukcji cyber-skuteczności elementów PPS (kolumna “Cybersec. Effectiveness of PPS element”) na ogólne prawdopodobieństwo detekcji i końcowy wynik oceny. Źródło: opracowanie własne na podstawie [41].

Analiza przedstawiona na rysunku 2.1 wskazuje na istotną różnicę w ocenie skuteczności systemu pomiędzy podejściem klasycznym (89,88%) a uwzględniającym zagrożenia cybernetyczne (1,63%). Różnica ta wynika z modelowanej redukcji cyber-skuteczności poszczególnych elementów ESS do 10% wartości nominalnych w wyniku potencjalnych ataków na infrastrukturę ICT.

W przedstawionym scenariuszu każdy element systemu detekcji podlega potencjalnemu kompromitowaniu cybernetycznemu, co prowadzi do kumulatywnego efektu obniżenia skuteczności całego systemu ochrony. Wyniki te podkreślają konieczność opracowania metodologii oceny, które systematycznie uwzględniają zarówno zagrożenia fizyczne, jak i cybernetyczne w ramach zunifikowanego podejścia analitycznego. Propozycja takiej metodologii zostanie przedstawiona w dalszej części niniejszej rozprawy.

Klasyczny EASI nie przewiduje takich scenariuszy, co prowadzi do dramatycznego przeszacowania skuteczności systemu [41].

Brak uwzględnienia degradacji

Modele takie jak EASI przyjmują skuteczność komponentów jako stałą, ignorując:

– Degradację fizyczną komponentów (starzenie elektroniki, korozja)

– Degradację oprogramowania (brak aktualizacji, akumulacja błędów)

– Degradację proceduralną (rutynizacja, erozja procedur)

– Degradację względem ewoluującego zagrożenia (nowe techniki ataku)

Wpływ czynnika degradacji na ocenę skuteczności systemów ochrony fizycznej ilustrują rysunki 2.2 i 2.3. Porównanie przedstawia ocenę tej samej ścieżki ataku metodą EASI w scenariuszu czysto fizycznym, z uwzględnieniem i bez uwzględnienia degradacji komponentów systemu.

Analiza porównawcza przedstawiona na rysunkach 2.2 i 2.3 wskazuje na redukcję ocenianej skuteczności systemu z 88,64% do 76,19% wyłącznie w wyniku uwzględnienia degradacji związanej z wiekiem poszczególnych komponentów. W drugim scenariuszu uwzględniono zróżnicowany wiek elementów ESS: czujnik światłowodowy (3 lata), kamery CCTV (7 lat), kontaktrony drzwiowe (5 lat), detektory PIR (8 lat) oraz kamery CCTV w funkcji sabotażu (7 lat).

Różnica 12,45 punktu procentowego między obiema ocenami ilustruje istotność uwzględnienia czynników degradacyjnych w procesie oceny skuteczności PPS. Wyniki te wskazują, że tradycyjne podejścia mogą prowadzić do przeszacowania rzeczywistej skuteczności systemów eksploatowanych przez dłuższy okres, co ma bezpośrednie implikacje dla planowania konserwacji i modernizacji infrastruktury ochronnej. Szczegółowy model degradacji zastosowany w niniejszej analizie zostanie przedstawiony w dalszej części rozprawy.

Badania pokazują spadek efektywności systemów nawet o 50% w ciągu 6-8 lat [6].

Ręczny charakter analiz

Klasyczne metody wymagają ręcznego modelowania sekwencji ataku, co skutkuje:

– Czasochłonnością (40-80 godzin dla średniego obiektu)

– Podatnością na błędy eksperckie

– Brakiem skalowalności

Response Relaybility	Mean time of response (sec.)	Force Time standard dev. (sec.)
0,95	250	90

Task #:	Adversary task description	Pd (probability of detection)	Detection Location (B/M/E)	mean delay	sdev	Normal values	cum PI
1	cut fence	90%	B	90	3	0,950385092	0,8125793
2	run	90%	B	12	3,6	0,784149763	0,0670448
3	open a door	90%	B	90	27	0,749086947	0,0064047
4	run	90%	B	10	3	0,421259148	0,0003602
5	open a door	90%	B	100	27	0,382797242	3,273E-05
6	sabotage of the asset	90%	B	120	36	0,089938638	7,69E-07
7		0%	0	0	0	0,002736602	0
8		0%	0	0	0	0,002736602	0
9		0%	0	0	0	0,002736602	0
10		0%	0	0	0	0,002736602	0
11		0%	0	0	0	0,002736602	0
12		0%	0	0	0	0,002736602	0
						Probability of interuption (Physical attack)	0,8864

Rysunek 2.2: Ocena skuteczności PPS metodą EASI dla czysto fizycznego wektora ataku bez uwzględnienia degradacji komponentów. Tabela przedstawia nominalne prawdopodobieństwa detekcji (Pd = 90%) dla wszystkich elementów systemu, skutkując prawdopodobieństwem przerwania ataku na poziomie 88,64%. Źródło: opracowanie własne na podstawie [6].

Response Relaybility	Mean time of response (sec.)	Force Time standard dev. (sec.)
0,95	250	90

Task #:	Adversary task description	Detection Element/s	Age of system	Mean of estimated efectivenes from degradation model	Probability of detection on the installation day	Pd (probability of detection on the day of evaluation)	Detection Location (B/M/E)	mean delay	sdev	Normal values	cum Pl
1	cut fence	Fiber-optic sensor	3	70%	90%	63%	B	90	3	0,950385092	0,5688055
2	run	CCTV camera	7	30%	90%	27%	B	12	3,6	0,784149763	0,0744197
3	open a door	Door contact	5	50%	90%	45%	B	90	27	0,749086947	0,0864954
4	run	PIR Detector	8	20%	90%	18%	B	10	3	0,421259148	0,0107012
5	open a door	Door contact	5	50%	90%	45%	B	100	27	0,382797242	0,0199345
6	sabotage of the asset	CCTV camera	7	30%	90%	27%	B	120	36	0,089938638	0,0015456
7						0%	0	0	0	0,002736602	0
8						0%	0	0	0	0,002736602	0
9						0%	0	0	0	0,002736602	0
10						0%	0	0	0	0,002736602	0
11						0%	0	0	0	0,002736602	0
12						0%	0	0	0	0,002736602	0
										Probability of interuption (Physical attack)	0,7619

Rysunek 2.3: Ocena skuteczności PPS metodą EASI z uwzględnieniem degradacji komponentów w czasie eksploatacji. Kolumna “Age of system” określa wiek poszczególnych elementów, a kolumna “Mean of estimated effectiveness from degradation model” pokazuje efektywność po uwzględnieniu procesu starzenia. Końcowe prawdopodobieństwo przerwania ataku wynosi 76,19%. Źródło: opracowanie własne na podstawie [6].

– Statycznością analiz

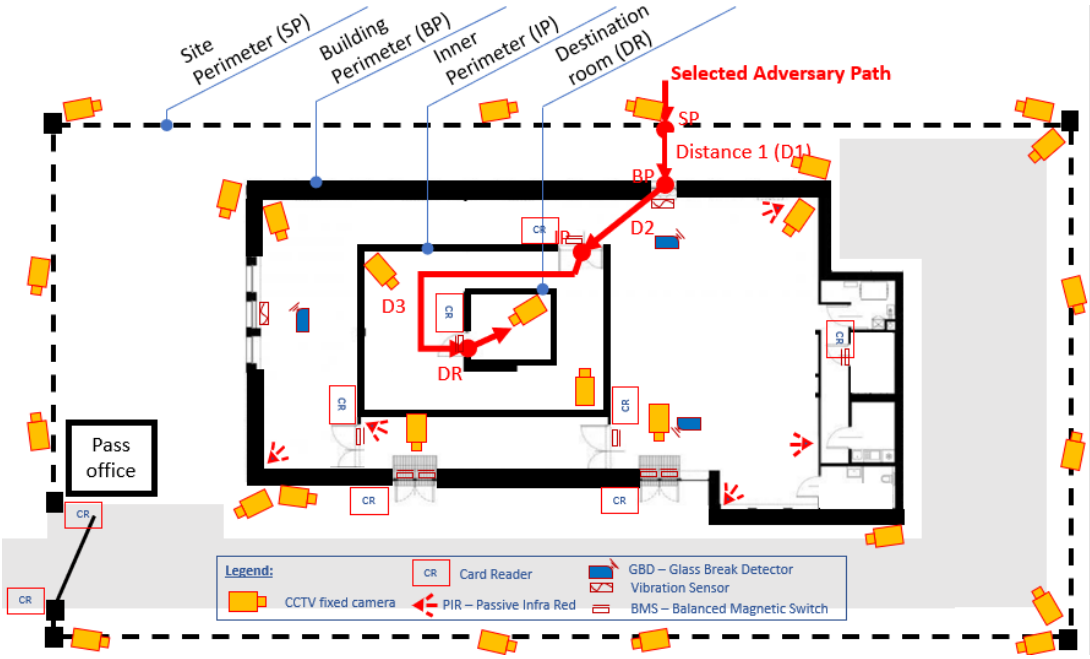
Ilustracją tego problemu jest przedstawiony na rysunku 2.4 przykład pojedynczej ścieżki ataku, opracowanej w ramach badań nad oceną cyberodporności systemów ochrony fizycznej [41]. Proces utworzenia tej ścieżki wymagał szczegółowej analizy eksperckiej uwzględniającej nie tylko topologię budynku, lecz również precyzyjne mapowanie wszystkich elementów ESS oraz ich wzajemnych relacji przestrzennych.

Analiza przedstawiona na rysunku 2.4 obejmuje jedną z wielu możliwych ścieżek infiltracji. W rzeczywistości, nawet dla tego stosunkowo prostego obiektu, liczba potencjalnych wariantów ścieżek ataku może sięgać dziesiątek tysięcy kombinacji, co czyni wyczerpującą analizę manualną praktycznie niewykonalną w akceptowalnym czasie i przy rozsądnych kosztach. Każda zmiana w konfiguracji systemu, wprowadzenie nowych elementów zabezpieczeń lub modyfikacja układu architektonicznego wymaga ponownej, czasochłonnej analizy eksperckiej.

Dodatkowo, ręczne podejście charakteryzuje się wysoką wrażliwością na czynnik ludzki - różni eksperci mogą identyfikować odmienne ścieżki jako krytyczne, co wprowadza element subiektywności do procesu oceny. Problem ten jest szczególnie istotny w kontekście infrastruktury krytycznej, gdzie spójność i powtarzalność analiz bezpieczeństwa mają kluczowe znaczenie dla zapewnienia odpowiedniego poziomu ochrony.

Koncepcja rozszerzenia

W odpowiedzi na lukę związaną z brakiem uwzględnienia cyberzagrożeń, zaproponowaliśmy praktyczną modyfikację modelu EASI polegającą na wprowadzeniu współczynnika cyberodporności (CR) dla



Rysunek 2.4: Przykład ręcznie utworzonej pojedynczej ścieżki ataku na chroniony obiekt. Czerwona linia przedstawia trasę przeciwnika od punktu wejścia na teren (SP - Site Perimeter) przez kolejne strefy ochrony: Distance 1 (D1), Building Perimeter (BP), Inner Perimeter (IP) do docelowego pomieszczenia (DR - Destination Room). Plan ilustruje rozmieszczenie elementów ESS: kamery CCTV (żółte prostokąty), detektory tłuczenia szkła GBD (niebieskie prostokąty), czujniki ruchu PIR (czerwone strzałki), czujniki wibracji oraz przełączniki magnetyczne BMS. Źródło: opracowanie własne na podstawie [41].

każdego cyfrowego komponentu systemu [41].

Modyfikacja polega na skorygowaniu prawdopodobieństwa detekcji:

$$P_{Di_{cyber}} = P_{Di} \cdot CR_i$$

gdzie:

– $P_{Di_{cyber}}$ - skorygowane prawdopodobieństwo detekcji z uwzględnieniem cyberzagrożeń

– P_{Di} - nominalne prawdopodobieństwo detekcji (przy założeniu braku cyberataków)

– CR_i - współczynnik CR i-tego komponentu (wartość od 0 do 1)

Fizyczne Systemy Ochrony (PPS) stanowią obowiązkowy element strategii ochrony narodowej infrastruktury krytycznej. Są one również powszechnie stosowane w wielu innych sektorach, w tym na rynku bankowym. Większość publikacji koncentrujących się na PPS odnosi się jednak do elektrowni jądrowych. Wydaje się oczywiste, że organizacja wydająca środki na wdrożenie PPS jest zainteresowana racjonalnym wykorzystaniem przydzielonego budżetu oraz skutecznością zainstalowanego rozwiązania.

Problem polega na tym, że EASI (wprowadzona w 1977 roku przez Bennetta) nie uwzględnia cyberataku jako części ataku fizycznego, podczas gdy obecnie stosowane PPS często wykorzystują sieci technologii informacyjno-komunikacyjnych (ICT). ICT dały nowe możliwości w zakresie funkcjonalności i skuteczności PPS, ale jednocześnie otworzyły je na nowe zagrożenia. Prowadzi to do wniosku, że jeśli jakakolwiek część funkcjonalności PPS może zostać przejęta lub zakłócona przez cyberprzestrzeń, może to znacząco wpłynąć na wyniki każdej „klasycznej” (wykorzystującej EASI) metody oceny. W związku z tym system oceniany jako skuteczny może nie być tak skuteczny, jak moglibyśmy oczekiwać.

EASI jako taka sprawdza prawdopodobieństwo zatrzymania przeciwnika przez siły reagowania, biorąc pod uwagę elementy takie jak prawdopodobieństwo wykrycia, czas opóźnienia (czas potrzebny na pokonanie bariery/warstwy ochronnej i odległość między nimi), czas reakcji i niezawodność komunikacji, aby obliczyć prawdopodobieństwo przerwania (P_i). Największą zaletą jest jej prostota. Jednak pierwotnie jest przeznaczona do oceny wybranej ścieżki przeciwnika, przy założeniu, że przeciwnik działa wyłącznie fizycznie.

Znaczenie czynnika cybernetycznego zostało zauważone przez specjalistów ds. bezpieczeństwa fizycznego. Jednak prawdopodobnie ze względu na specjalizację różnych dyscyplin bezpieczeństwa, wydaje się, że istnieje tendencja do utrzymywania go na bardzo wysokim poziomie ogólności i podjęto bardzo niewiele prób jego sformalizowania. Autorzy publikacji takich jak „Investigation of Cyber-Enabled Physical Attack Scenarios"[9] oraz „Physical Protection Systems and Cyber Security Component"[10] również doszli do wniosku, że interesariusze PPS nie posiadają niezbędnej wiedzy związanej z cyberatakami na PPS oraz że wymagana jest współpraca między bezpieczeństwem fizycznym a cybernetycznym. Oba wnioski zostały sformułowane w 2015 roku i wydaje się, że niewiele się w tym zakresie zmieniło. Stosunkowo często możemy znaleźć specjalistów PPS podnoszących temat znaczenia cyberbezpieczeństwa dla PPS, ale niewiele szczegółowych informacji na temat tego, jak nim zarządzać.

Określanie współczynnika cyberodporności

Ciekawą i obiecującą ideę oceny skuteczności PPS z uwzględnieniem cyberataków wprowadził J. Depoy i współautorzy [11]. Zgodnie z ich propozycją, połączenie EASI z BATLE (Brief Adversary Threat Loss Estimator) [44] powinno być wykorzystywane do oceny skuteczności PPS przeciwko atakowi fizycznemu. Jednak w przypadku ataku fizycznego wspomaganego cybernetycznie, ogólna skuteczność PPS PE_{cp} powinna być obliczana jako kombinacja prawdopodobieństwa sukcesu ataku fizycznego $(1-PE_{pz})$ i prawdopodobieństwa sukcesu cyberataku $(1 - PE_c)$. Poniższe równanie to przedstawia:

$$PE_{cp} = 1 - (1 - PE_c) \cdot (1 - PE_{pz})$$

Autorzy słusznie zauważyli, że wydajność elementów sterowanych cybernetycznie PPS wynosi zero, jeśli zostaną one pokonane przez cyberprzestrzeń. Jednak przykład, który przedstawili, koncentruje się na zmniejszeniu czasu opóźnienia (czasu potrzebnego na pokonanie elementu ochronnego) do zera, podczas gdy równie ważny, jeśli nie ważniejszy, jest wpływ na współczynnik detekcji w EASI. Innymi słowy, jeśli funkcja detekcji PPS zostanie wyłączona przez cyberatak, czas opóźnienia może nie być tak ważny, jak fakt, że czas reakcji nie zacznie być liczony.

Podobne podejście przedstawili Snell i Rivers [12]. Nawet równanie wygląda podobnie do przedstawionego przez J. Depoy'a i współautorów, więc skuteczność systemu byłaby obliczana następująco:

$$PE_{Total} = 1 - (1 - PDS) \cdot (1 - PEO)$$

W tym przypadku PDS to prawdopodobieństwo detekcji, a PEO to zakładana ocena sukcesu dowolnego ataku fizycznego i cybernetycznego. Główna różnica w porównaniu z wcześniej wspomnianą propozycją polega na tym, że tutaj autorzy wykorzystali ASD (adversary sequence diagram) i DBT (Design Basis Threat) [45]. Oznacza to, że ich podejście jest bardziej oparte na scenariuszach i uwzględnia zestaw potencjalnych ścieżek przeciwnika zamiast jednej. Jednak ich propozycja jest bardziej kierunkiem niż kompletną metodologią i w większości opiera się na wiedzy eksperckiej.

Nowsze prace, takie jak autorstwa Guo, Yan i Wang [13], nadal pokazują znaczenie cyberbezpieczeństwa dla PPS. Jednak koncentrują się bardziej na ryzyku niż na metodach oceny skuteczności PPS w kontekście ataków wspomaganych cybernetycznie. Podobnie Bowen i współautorzy [14] podkreślili ważną rolę PPS i wykorzystali nowoczesne narzędzia do modelowania i wizualizacji ścieżki przeciwnika. Jednak w ich podejściu brakuje czynnika potencjalnego cyberataku na infrastrukturę bezpieczeństwa fizycznego. Również Andiwijayakusuma i współautorzy [15] szukają poprawy EASI w odniesieniu do analizy wielościeżkowej bez uwzględnienia czynnika cybernetycznego w swojej metodzie.

Ze względu na powyższe uważamy, że stosunkowo duża luka w cyberbezpieczeństwie w ocenie skuteczności PPS nadal wymaga wypełnienia.

Przykład zastosowania

Zgodnie z wytycznymi Mary Lynn Garcia [7], należało stworzyć „sekwencję działań przeciwnika"(ang. AAS - Adversary Actions Sequence) i wybrać ścieżkę fizyczną odpowiadającą ASS. W następ-

nym kroku ASS i ścieżka powinny zostać zwizualizowane. Aby wykonać to zadanie, należało narysować ASS na planie hipotetycznego obiektu.

W następnym kroku narzędzie oparte na Excelu musiało zostać wypełnione zebranymi informacjami, takimi jak:

- Krok sekwencji działań przeciwnika
- Prawdopodobieństwo detekcji dla każdego elementu detekcji
- Średnia i odchylenie standardowe czasu każdego zadania
- Typ detekcji (przed (B), w środku (M) lub na końcu (E) opóźnienia)
- Niezawodność komunikacji

W wyniku EASI zwróciło prawdopodobieństwo przerwania sekwencji przeciwnika przez siły reagowania.

Nowe podejście - kierunek i heurystyka

Bazując na fakcie, że zarówno J. Depoy i współautorzy [11], jak i Snell i Rivers [12] mieli w pewnym momencie rację w swoich podejściach, doszliśmy do wniosku, że jako „łatwą poprawkę"możemy wykorzystać następujące wyniki naszej wstępnej analizy:

Poziom cyberodporności komponentów PPS może się różnić. Dla naszych celów, na podstawie doświadczenia, wstępnie arbitralnie przyjęliśmy trzy poziomy „skuteczności cyberbezpieczeństwa”:

- 1 - Wysoki równy 99%
- 2 - Umiarkowany równy 50%
- 3 - Niski równy 10%

Cyberatak może mieć negatywny wpływ na skuteczność detekcji detektorów, na czas opóźnienia barier fizycznych lub na oba te elementy. Cyberatak może mieć negatywny wpływ na niezawodność sił reagowania.

Biorąc pod uwagę powyższe, zmodyfikowaliśmy narzędzie EASI oparte na Excelu. W zmodyfikowanej wersji całkowite prawdopodobieństwo detekcji kroku przeciwnika jest obliczane jako iloczyn prawdopodobieństwa detekcji i skuteczności cyberbezpieczeństwa elementów PPS:

$$P_{DT} = P_d \cdot P_{ce}$$

Czas opóźnienia (jeśli wpływa na niego potencjalny cyberatak) jest redukowany do stałej wartości dziesięciu sekund w przypadku „niskiego"poziomu skuteczności cyberbezpieczeństwa. W przypadku „umiarkowanej"skuteczności cyberbezpieczeństwa czas opóźnienia jest redukowany o 50% jego pierwotnej wartości i o 1% dla skuteczności ocenianej jako „wysoki"poziom. Wybraliśmy czas dziesięciu sekund ze względu na założenie, że nawet jeśli element sterowany cybernetycznie zostanie zdalnie przejęty, czas

pokonania określonej bariery nigdy nie zostanie zredukowany do zera. Na przykład, jeśli przeciwnik musi przejść przez drzwi z elektronicznym zamkiem sterowanym cybernetycznie, aby je otworzyć, będzie musiał albo sam połączyć się z systemem, albo skontaktować się z kimś, kto otworzy je dla niego. Tak czy inaczej, zajmie to trochę czasu i bez wątpienia nie będzie to zero. Czas dziesięciu sekund wydaje się rozsądny, aby pokryć dobrze zaprojektowane kroki wymagane do otwarcia i przejścia przez barierę sterowaną cybernetycznie.

Wypełniliśmy zmodyfikowane narzędzie i w rezultacie otrzymaliśmy potwierdzenie naszego wstępnego założenia, że wynik oceny skuteczności PPS bez uwzględnienia działań przeciwnika z cyberprzestrzeni może być doskonały. Jednocześnie, jeśli cyberatak jest uwzględniony, końcowa ocena może być znacząco różna. W naszym przypadku ocena uwzględniająca wyłącznie fizyczne zadania przeciwnika osiągnęła wynik 89,88% prawdopodobieństwa, że ścieżka przeciwnika zostanie zakłócona, i tylko 1,63% prawdopodobieństwa w przypadku uwzględnienia cyberataku na PPS jako części ataku fizycznego.

Zmodyfikowaliśmy oryginalną formułę EASI, uwzględniając wpływ cyberbezpieczeństwa na prawdopodobieństwo detekcji (równanie 2.2) i czas opóźnienia. Czas opóźnienia w konsekwencji wpływa na prawdopodobieństwo przybycia sił reagowania na miejsce zdarzenia przed zakończeniem sekwencji przeciwnika $P(R|A)$. W ten sposób kumulatywne prawdopodobieństwo przerwania przeciwnika $(P(I))$ jest obliczane następująco:

$$P(I) = (P_{DT1}) \cdot (P_{C1}) \cdot (P(R|A_1)) + \sum_{i=2}^{n} \left[(P(R|A_i)) \cdot (P_{Ci}) \cdot (P_{DTi}) \cdot \prod_{j=1}^{i-1} (1 - (P_{DTj})) \right]$$

gdzie: P_{DTi} to prawdopodobieństwo detekcji podczas i-tej akcji przeciwnika, która musi być podjęta na i między warstwami bezpieczeństwa, obliczane zgodnie z równaniem (2.2); P_{Ci} to prawdopodobieństwo komunikacji z siłami reagowania; $P(R|A_i)$ to warunkowe prawdopodobieństwo przybycia sił reagowania i jest równe $P(X > 0)$, gdzie X jest równe pozostałemu czasowi (TR) minus pozostały czas reakcji (RTF).

Wyniki analizy

Przeprowadzona analiza hipotetycznej ścieżki ataku wykazała dramatyczną różnicę między oceną klasyczną a oceną uwzględniającą cyberzagrożenia. Dla analizowanego przypadku przyjęto następujące założenia:

Wyniki pokazują, że klasyczna ocena EASI wskazywała prawdopodobieństwo przerwania ataku na poziomie 89,88%, podczas gdy po uwzględnieniu współczynników cyberodporności wartość ta spadła do zaledwie 1,63%.

Poprawa EASI poprzez dodanie aspektu cyberbezpieczeństwa do głównej formuły pozwoliła nam znaleźć ten etap jako dobry punkt wyjścia do rozwiązania niektórych innych wyzwań związanych z EASI. Pierwszym z nich jest dobrze znany fakt, że EASI sprawdza pojedynczą ścieżkę. Zostało to częściowo rozwiązane w przeszłości przez niektóre inne metodologie (np. [15]) oceny PPS. Jednak procesy tworzenia sekwencji przeciwnika i wyboru ścieżki przeciwnika są nadal ręczne lub półautomatyczne. Oba

Element	P_d	**Poziom CR (cyberodporności)**	P_{ce}
Ogrodzenie zewnętrzne	90%	-	100%
Czujniki ruchu	90%	Umiarkowany	50%
Kamery CCTV	90%	Niski	10%
Kontrola dostępu (czytniki RFID)	90%	Niski	10%
Czujniki otwarcia	90%	Umiarkowany	50%
System alarmowy	90%	Niski	10%

Tabela 2.1: Parametry elementów PPS w analizowanej ścieżce ataku

procesy są zwykle prowadzone przez ekspertów, więc istnieje ryzyko, że niektóre podatne ścieżki mogą zostać pominięte.

Uważamy, że możemy wykorzystać wcześniej opisaną formułę (2.2) i sprawdzić wszystkie możliwe ścieżki przeciwnika w połączeniu z algorytmami przeszukiwania grafów i wyszukiwania ścieżek. Jesteśmy nadal we wczesnej fazie testowania, ale na tym etapie znaleźliśmy następującą (podzieloną na kroki) metodę jako obiecującą:

1. Dodanie węzłów do rysunku obiektu zawierającego elementy PPS. Węzły powinny reprezentować wszystkie potencjalnie atakowane bariery fizyczne. W przypadku barier takich jak ściany, ogrodzenia itp. powinny być one podzielone na kilka węzłów w zależności od rozmiaru i lokalizacji konkretnej części bariery.

2. Wykorzystując wcześniej utworzone węzły, stworzenie grafu wszystkich potencjalnych ścieżek przeciwnika.

3. Określenie lub obliczenie (w zależności od typu parametru) wartości parametrów dla każdego węzła i krawędzi. Początkowo powinny to być parametry takie jak:

   – Prawdopodobieństwo detekcji P_d

   – Skuteczność cyberbezpieczeństwa elementów PPS P_{ce}

   – Czas opóźnienia T_{dl}

   – Pozostały czas reakcji do określonego punktu na grafie T_{rr}

4. Użycie funkcji heurystycznej $f(P_d, P_{ce}, T_{dl}, T_{rr})$ i jednego z algorytmów przeszukiwania grafów i wyszukiwania ścieżek do przewidywania ścieżek, które są najbardziej podatne.

5. Sprawdzenie, czy istnieją węzły wspólne dla obliczonych podatnych ścieżek.

6. Wzmocnienie odkrytych słabych punktów, aż prawdopodobieństwo skutecznej interwencji sił reagowania spełni oczekiwania organizacji.

Wnioski z rozszerzenia

W wyniku naszej pracy zaproponowaliśmy modyfikację podejścia EASI do oceny Fizycznych Systemów Ochrony. Proponowana metoda obejmuje potencjalne zagrożenia wykorzystujące cyberatak przeciwko komponentom PPS jako część ataku fizycznego. Udowodniliśmy, że cyberataki mogą znacząco zmniejszyć wynik standardowej oceny EASI PPS.

Zaproponowaliśmy również kierunek dalszego rozwoju bardziej kompleksowego podejścia wykorzystującego funkcję heurystyczną i algorytmy przeszukiwania grafów i wyszukiwania ścieżek. Na obecnym etapie nadal szukamy najbardziej efektywnej funkcji heurystycznej. Jednak już zmodyfikowana formuła EASI daje obiecujące wyniki.

Innym aspektem wymagającym dalszych badań jest skuteczność mechanizmów cyberbezpieczeństwa chroniących elementy PPS. Jak wyjaśniono, podzieliliśmy ją na trzy poziomy (wysoki, umiarkowany i niski) jako podejście początkowe. Ten podział jest konsekwencją wstępnych wyników testów penetracyjnych, które przeprowadzamy w ramach innych badań będących w toku.

Kluczowym wnioskiem jest to, że organizacje opierające się wyłącznie na tradycyjnych metodach oceny mogą funkcjonować w fałszywym poczuciu bezpieczeństwa. Różnica wynosząca ponad 88 punktów procentowych między oceną klasyczną a oceną uwzględniającą cyberzagrożenia jednoznacznie wskazuje na konieczność uwzględnienia tego aspektu w każdej rzetelnej ocenie skuteczności PPS. Wprowadzenie współczynnika cyberodporności (CR) do klasycznej metodologii EASI stanowi niezbędne dostosowanie do realiów współczesnych systemów ochrony fizycznej, gdzie praktycznie każdy komponent zawiera elementy cyfrowe podatne na ataki z cyberprzestrzeni.

2.3. Model probabilistyczny degradacji systemów elektronicznych

Koncepcja modelu degradacji

Jak już podkreślano, brak uwzględnienia procesu degradacji elementów systemu jest kolejną istotną luką w aktualnych metodologiach oceny skuteczności PPS. Klasyczne podejścia, takie jak EASI, zakładają stałą skuteczność komponentów w czasie, co jest założeniem dalekim od rzeczywistości operacyjnej systemów bezpieczeństwa [6].

W kontekście Cyber-Fizycznych Systemów Ochrony (CPPS) degradację definiujemy jako negatywne zmiany w skuteczności elementów, wynikające z eksploatacji (np. obniżanie parametrów wraz z wiekiem) oraz zmian środowiskowych (np. pojawienie się nowych narzędzi ataku). Jako ilustrację problemu można przytoczyć przykład ochrony banków, gdzie ryzyko napadu znacząco spadło, a wzrosło ryzyko ataków na bankomaty (ATM). Historycznie dominowały ataki fizyczne (kradzież kluczy, wiercenie, kradzież kaset z pieniędzmi). Natomiast współcześnie, z uwagi na zmianę w wiedzy i świadomości ryzyka u przeciwnika, w obszarze fizycznym dominują ataki z użyciem środków wybuchowych (przeważnie gaz ziemny) oraz ataki cyber-fizyczne, co znacząco zwiększyło ryzyko i koszty zabezpieczeń, które należy dostosować do zmiany trendu w materializacji ryzyk.

Model probabilistyczny degradacji CPPS został opracowany w celu kwantyfikacji wpływu pięciu kluczowych czynników wpływających na efektywność systemów w czasie:

1. **AG (Aging)** - starzenie się komponentów
2. **KG (Knowledge Growth)** - wzrost wiedzy i umiejętności przeciwnika
3. **TI (Technology Improvement)** - postęp technologiczny narzędzi ataku
4. **VI (Vulnerability Introduction)** - pojawianie się nowych podatności
5. **IE (Improvement Effects)** - efekty działań konserwacyjnych

Formalizacja matematyczna modelu

Efektywność systemu w miesiącu i definiowana jest jako:

$$SE_i = \max(100 - X_i, 0)$$

gdzie X_i to skumulowany wpływ degradacyjny obliczany jako:

$$X_i = \sum_{j=1}^{i} (AG_j + KG_j + TI_j + VI_j - IE_j)$$

Zastosowanie funkcji $\max()$ w modelu $SE_i = \max(100 - X_i, 0)$ zapewnia, że efektywność systemu nigdy nie przyjmuje wartości ujemnych. Formalnie interpretujemy $SE_i = 0$ jako całkowitą utratę zdolności ochronnych CPPS, co odpowiada sytuacji, w której dalsza degradacja nie ma praktycznego znaczenia (system już jest nieskuteczny).

Przyjęte podejście modelowe znajduje uzasadnienie w najnowszych badaniach AlMuhaini [46], który wprowadził nowatorski model kwantyfikacji odporności dla cyber-fizycznych systemów energetycznych oparty na "trapezie odporności". Model ten uwzględnia różne fazy degradacji zarówno w domenie cybernetycznej, jak i fizycznej podczas ekstremalnych zdarzeń naturalnych lub antropogenicznych, wykorzystując symulacje Monte Carlo dla oszacowania poziomów odporności węzłów lokalnych i całych systemów.

Parametryzacja rozkładów probabilistycznych

W wyniku analizy współczesnego środowiska bezpieczeństwa, obejmującego dynamicznie zmieniające się zagrożenia technologiczne, wzrost kompetencji adwersarzy oraz praktyki utrzymania działania systemów, każdy z pięciu czynników modelowany jest odpowiednim rozkładem probabilistycznym. Wartości parametrów zostały dobrane w oparciu o ramy czasowe sugerowane przez ekspertów (5-10 lat cyklu życia ESS) oraz wyniki badań literaturowych.

Starzenie się komponentów (AG) modelowane jest rozkładem normalnym $N(0.5, 0.15)$, co odzwierciedla średni spadek sprawności o około 50 punktów procentowych w dekadzie (rozrzut ±15 pp). Starzenie jest efektem sumy wielu drobnych, w dużej mierze niezależnych mikrozjawisk (zużycie materiałów, dryf parametrów), co uzasadnia wybór rozkładu normalnego. Wartości te są zgodne ze wskazaniami literaturowymi dotyczącymi degradacji elementów elektronicznych [16, 17].

Wzrost wiedzy przeciwnika (KG) opisany jest rozkładem normalnym $N(0.1, 0.03)$, co odpowiada około 10% rocznemu przyrostowi skuteczności ataków dzięki lepszemu planowaniu i dostępowi do wiedzy. Akumulacja wiedzy i umiejętności następuje przyrostowo, z naturalną zmiennością w tempie i jakości pozyskiwanych danych [11].

Postęp technologiczny (TI) modelowany jest funkcją logarytmiczną $TI_j = \mu \cdot \log(j) + \epsilon$, gdzie $\mu = 0.7$, $\epsilon \sim N(0, 0.21)$. W początkowych fazach rozwoju narzędzi obserwuje się szybkie przyrosty efektywności, które następnie maleją (diminishing returns). Funkcja logarytmiczna dobrze oddaje tę dynamikę, co potwierdzają dyskusje i przykłady trendów technologicznych w kontekście PPS/CPPS [13].

Wpływ podatności (VI) reprezentowany jest procesem Bernoulliego z parametrem $p = 0.2$ i jednorazowym spadkiem efektywności o 10 pp przy wystąpieniu. Modeluje to zdarzenia rzadkie, skokowe, o trudnej do przewidzenia częstości w krótkich horyzontach (np. zero-day). Historyczna częstość wynosi około 1 na 5 miesięcy [15].

Efekty kwartalnych usprawnień (IE) opisane są rozkładem normalnym $N(0.5, 0.15)$, modelującym prewencyjne działania konserwacyjne (testy, wymiana części) co 3 miesiące, redukujące skumulowany spadek efektywności o średnio 50 pp z odchyleniem ±15 pp. Łączny efekt wielu czynności serwisowych z naturalną zmiennością zakresu i jakości prac uzasadnia modelowanie sumarycznego efektu rozkładem normalnym.

Integracja modelu degradacji z metodologią EASI

Włączenie modelu degradacji do EASI odbywa się przez modyfikację prawdopodobieństw detekcji każdego elementu na ścieżce ataku:

$$P_{Di_{deg}} = P_{Di} \cdot \frac{SE_i}{100}$$

gdzie $P_{Di_{deg}}$ to prawdopodobieństwo detekcji z uwzględnieniem degradacji, P_{Di} to nominalne prawdopodobieństwo detekcji przy założeniu nowego, w pełni sprawnego elementu, a SE_i to skuteczność elementu po i miesiącach eksploatacji zgodnie z modelem degradacji.

Ta modyfikacja bezpośrednio wpływa na obliczenia prawdopodobieństwa przerwania ataku w formule EASI. Degradacja elementów prowadzi do obniżenia ich zdolności detekcyjnych, co w konsekwencji zmniejsza całkowitą skuteczność systemu ochrony.

Środowisko obliczeniowe i implementacja

Model został zaimplementowany w języku Python, co umożliwiło wykorzystanie nowoczesnych bibliotek do obliczeń numerycznych i wizualizacji wyników. W szczególności wykorzystaliśmy:

- **numpy** - podstawowa biblioteka do przechowywania i wektoryzowanych obliczeń na wielowymiarowych tablicach danych, zapewniająca wydajność zbliżoną do kodu C/C++
- **scipy.stats** - moduł do statystycznego modelowania rozkładów prawdopodobieństwa, wykorzystywany do generowania próbek z rozkładów normalnego, Bernoulliego oraz obliczania funkcji gęstości i dystrybuant

– **matplotlib** - biblioteka do tworzenia wizualizacji: wykresów liniowych przebiegów degradacji, histogramów rozkładów parametrów oraz linii trendu

W symulacjach przyjęto $N = 10,000$ niezależnych przebiegów. Taka liczba zapewnia, że oszacowania średniej efektywności i odchyleń standardowych są stabilne, przy akceptowalnym czasie obliczeń. Zgodnie z prawem wielkich liczb błąd standardowy estymowanej średniej maleje w tempie $1/\sqrt{N}$, co dla $N = 10,000$ skutkuje rzędem wielkości około 1%, dając wiarygodne wyniki.

Na początku każdej symulacji zostało ustawione ziarno losowości (seed), dzięki czemu sekwencja liczb pseudolosowych jest powtarzalna, co jest niezbędne do rzetelnej walidacji modelu oraz porównywania różnych scenariuszy testowych.

Przykład zastosowania - studium przypadku

Aby zilustrować zastosowanie modelu degradacji CPPS, rozważono typową ścieżkę ataku na obiekt chroniony, którego charakterystyka składa się z czterech kolejnych warstw ochronnych:

– Ochrona obwodowa - sensometryczny przewód światłowodowy monitorujący linię ogrodzenia

– Nadzór wizyjny - kamery CCTV nadzorujące obszar chroniony

– Kontrola dostępu - czytniki transponderów RFID oraz czujniki otwarcia (kontaktrony)

– System Sygnalizacji Włamania - pasywne czujki podczerwieni (PIR) wykrywające ruch wewnątrz obiektu

Każdemu z tych elementów przypisaliśmy przykładowy "wiek": przewód sensometryczny - 2 lata, kamery CCTV - 7 lat, czujniki otwarcia drzwi - 5 lat, czujki PIR - 10 lat.

W oparciu o model obliczono dla każdego elementu wartość efektywności SE_i odpowiadającą aktualnemu wiekowi urządzenia:

Element	**Wiek**	SE_i	P_{Di}	$P_{Di_{deg}}$
Przewód sensometryczny	2 lata	91%	0.90	0.819
Kamery CCTV	7 lat	45%	0.90	0.405
Czytniki RFID	5 lat	68%	0.90	0.612
Czujki PIR	10 lat	23%	0.90	0.207

Tabela 2.2: Wpływ degradacji na prawdopodobieństwo detekcji poszczególnych elementów

Następnie wykonaliśmy dwie oceny efektywności systemu CPPS atakiem wybraną wcześniej ścieżką. W klasycznym podejściu EASI zakładamy nominalne prawdopodobieństwa detekcji (P_{Di}) i transmisji alarmu (P_{Ci}) bez uwzględniania upływu czasu ani degradacji. Kumulatywne prawdopodobieństwo przerwania ataku $P(I)$ wyliczane jest według standardowego równania EASI. W zmodyfikowanej ocenie każdy nominalny współczynnik detekcji P_{Di} został przemnożony przez SE_i, czyli efektywność danej warstwy wynikającą z jej wieku i degradacji. Dzięki temu prawdopodobieństwo przerwania ataku maleje

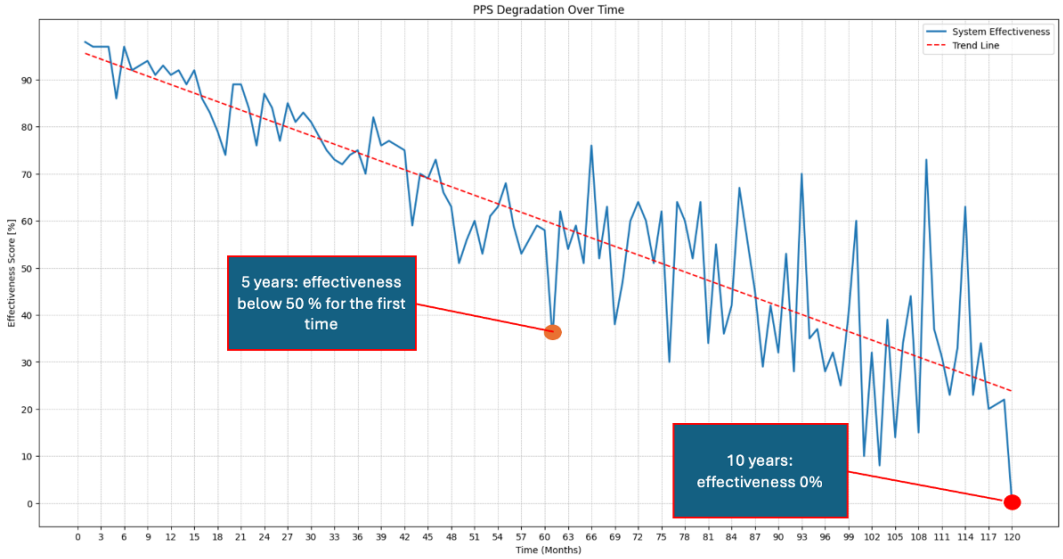
wraz ze starzeniem się elementów ochronnych. Jak zostało to wcześniej zilustrowane w sekcji 2.3 (patrz rysunki 2.2 i 2.3, klasyczna ocena EASI wskazała $P(I) = 88.64\%$. Ocena z uwzględnieniem degradacji: $P(I) = 76.19\%$. Różnica wynosi 12.45 punktów procentowych.

Wyniki symulacji Monte Carlo

Przeprowadzone symulacje Monte Carlo z 10,000 przebiegów modelujących degradację systemu w horyzoncie 120 miesięcy ujawniły charakterystyczny przebieg degradacji: system utrzymuje wysoką sprawność w pierwszych latach, po czym następuje przyspieszone obniżenie efektywności, z okresowymi, niewielkimi odbiciami związanymi z konserwacją.

Kluczowe obserwacje z symulacji:

Moment przekroczenia 50%: Chociaż pojedyncze realizacje modelu (niebieska linia) czasem schodzą poniżej 50% już po 60 miesiącach, średnia tendencja (linia trendu) wskazuje, że poziom 50% zostaje przekroczony dopiero około 75-80 miesiąca (6.5-6.7 roku). Oznacza to, że w rozważanym scenariuszu typowy cykl wymiany CPPS krótszy niż 7 lat byłby uzasadniony, jeżeli chcemy utrzymać skuteczność powyżej połowy nominalnej (patrz rysunek 2.5)



Rysunek 2.5: Wykres efektu symulacji degradacji systemów bezpieczeństwa w czasie. Źródło: opracowanie własne na podstawie [6].

Dominujące czynniki degradacji: Analiza poszczególnych składników X_i pokazuje wyraźne, nierówne udziały. Starzenie się (AG_j) rośnie liniowo od 0 do około 60-80 punktów procentowych w dekadzie i stanowi największy udział w skumulowanym spadku. Wzrost wiedzy przeciwnika (KG_j) narasta stopniowo od 0 do około 10-15 pp w ciągu dziesięciu lat. Udoskonalanie narzędzi (TI_j) przyspiesza w drugiej połowie okresu i sięga do ok. 8-10 pp. Podatności (VI_j) dają sporadyczne skoki spadku do 5-10 pp, ale są rzadkie (średnio jeden na pięć miesięcy). Efekty konserwacji (IE_j) pojawiają się kwartalnie i mają niewielką wartość - rzędu 1-2 pp regeneracji na cykl, co jest niewystarczające, by skompensować skumulowane straty.

Na podstawie porównania obu ocen widać, że uwzględnienie degradacji obniża znaczniki skuteczności CPPS o dodatkowe 15-25 punktów procentowych już w 5. roku użytkowania. W klasycznym podejściu (bez degradacji) CPPS wygląda na relatywnie znacznie bardziej odporne, co może prowadzić do błędnych decyzji, w tym o zbyt rzadkiej wymianie sprzętu.

Implikacje praktyczne modelu degradacji

W świetle wyników symulacji słusznym wydaje się rozważenie wprowadzenia ciągłego monitorowania efektywności i predykcji degradacji oraz adaptacyjnych harmonogramów konserwacji, które uwzględniają rzeczywiste tempo zużycia i zmienność zagrożeń. Model wymaga kalibracji na danych z rzeczywistych instalacji CPPS, zwłaszcza w odniesieniu do kwartalnych efektów konserwacji i częstotliwości występowania krytycznych podatności. Dopiero wówczas będzie możliwe opracowanie precyzyjnych, ekonomicznie zoptymalizowanych harmonogramów konserwacji i wymiany komponentów. Integracja wskaźnika efektywności SE_i z klasycznym modelem EASI pozwoliła na uzyskanie bardziej realistycznych i konserwatywnych estymat prawdopodobieństwa przerwania ataku, obniżając je o dodatkowe 15-25 punktów procentowych już po pięciu latach eksploatacji. Taka modyfikacja wydaje się być istotnym parametrem z punktu widzenia planowania budżetowego i zabezpieczenia ciągłości działań ochronnych.

Na tym etapie wydaje się, że zaproponowany model stanowi dobrą podstawę do przejścia od statycznych harmonogramów wymiany narzędzi ochronnych do dynamicznych, opartych na danych i analizie ryzyka procesów utrzymaniowych. Konieczne jest jednak jego dalsze rozwijanie i walidacja na rzeczywistych danych operacyjnych z instalacji CPPS.

2.4. Łączny wpływ cyberzagrożeń i degradacji

Naturalną konsekwencją odrębnych analiz wpływu cyberzagrożeń i degradacji jest ich łączne ujęcie w spójnym modelu oceny, wzmacniającym klasyczną metodologię EASI. W tej sekcji integrujemy oba wymiary na poziomie elementów ścieżki ataku: bazowe prawdopodobieństwo detekcji korygujemy o wpływ degradacji w czasie oraz oddziaływań cybernetycznych, przyjmując ich ujęcie multiplikatywne (iloczynowe). Tak skonstruowana rama zachowuje kompatybilność z aparatem obliczeniowym EASI, a jednocześnie pozwala uchwycić dynamiczny i kontekstowy charakter systemu.

Model zintegrowany

Połączenie obu rozszerzeń prowadzi do następującej kompleksowej formuły:

$$P_{Di_{total}} = P_{Di} \cdot CR_i \cdot \frac{SE_i}{100}$$

Analiza przypadku ekstremalnego

Dla systemu z umiarkowaną cyberodpornością (średnie $CR = 0.5$) po 5 latach eksploatacji (średnie $SE = 70\%$):

– Klasyczne EASI: $P(I) = 89.88\%$

– Tylko cyberzagrożenia: $P(I) = 1.63\%$

– Tylko degradacja: $P(I) = 76.19\%$

– Model zintegrowany: $P(I) < 1\%$

To względnie duże obniżenie skuteczności podkreśla celowość uwzględnienia obu czynników w ocenie CPPS.

2.5. Podsumowanie

W niniejszym rozdziale wykazaliśmy, że klasyczna metodologia EASI, mimo szerokiego zastosowania, nie odpowiada w pełni realiom współczesnego środowiska zagrożeń. Kluczowe ograniczenia, takie jak nieuwzględnianie cyberzagrożeń, pomijanie zjawisk degradacyjnych oraz manualny charakter analiz, mogą prowadzić do znacznego przeszacowania skuteczności systemów ochrony.

Zaproponowane rozszerzenie metodologii EASI wydaje się stanowić dobrą podstawę do wdrożenia bardziej zautomatyzowanego i spójnego procesu oceny, który adresuje wskazane luki. Poszczególne komponenty tego podejścia zostaną przedstawione w kolejnym rozdziale.

3. Detekcja elementów budynku i infrastruktury bezpieczeństwa

Jak wykazano wcześniej, skala problemu oceny bezpieczeństwa rośnie wykładniczo — liczba potencjalnych ścieżek ataku w realistycznym obiekcie może być astronomiczna (np. dla 36 węzłów startowych i maksymalnie 40 kroków: 4.6×10^{25} kombinacji), co czyni ręczną analizę niepełną, subiektywną i mało powtarzalną. Aby zminimalizować te ograniczenia, niniejszy rozdział koncentruje się wyłącznie na pierwszym, fundamentalnym etapie automatyzacji: **detekcji elementów budynku i infrastruktury bezpieczeństwa** na planach 2D.

Prezentujemy: (i) konfigurację i wyniki detekcji elementów architektonicznych (YOLOv8), (ii) postprocessing i rafinację detekcji z użyciem IoSA i walidacji geometrycznej, oraz (iii) interpretację symboli systemów bezpieczeństwa z wykorzystaniem *COG*. Kwestie generowania grafów i heurystycznej oceny ścieżek są przedstawione w odrębnych rozdziałach.

3.1. Ewolucja detekcji obiektów w kontekście bezpieczeństwa

Od klasycznych metod do głębokiego uczenia

Zastosowanie wizji komputerowej w bezpieczeństwie przeszło znaczącą ewolucję od prostych systemów detekcji ruchu, przez zaawansowane algorytmy śledzenia, aż po współczesne systemy oparte na głębokim uczeniu. Kluczowe etapy tej ewolucji obejmują [28, 29, 31]:

Era klasycznych deskryptorów (2000-2012): Wczesne systemy bezpieczeństwa wykorzystywały deskryptory cech takie jak SIFT [28] i SURF, które umożliwiały rozpoznawanie obiektów na podstawie charakterystycznych punktów kluczowych. Metody te, choć innowacyjne, charakteryzowały się ograniczoną odpornością na zmienność oświetlenia i deformacje geometryczne.

Rewolucja Konwolucyjnych Sieci Neuronowych - KSN (2012-2014): Przełomowe znaczenie miało wprowadzenie AlexNet [29], które zademonstrowało potencjał KSN w klasyfikacji obrazów. Kolejne architektury, VGG [30] i ResNet [31], ustanowiły nowe standardy dokładności, jednak ich aplikacja w detekcji obiektów wymagała dodatkowych rozwiązań.

Rozwój detektorów dwuetapowych (2014-2016): Regionalnych Konwolucyjnych Sieci Neuronowych (ang. R-CNN) [32] zrewolucjonizował detekcję obiektów, wprowadzając dwufazową generację propozycji regionów i klasyfikację. "Szybkie"Regionalne Konwolucyjne Sieci Neuronowe (ang. Fast R-

CNN) i "Szybsze"Regionalne Konwolucyjne Sieci Neuronowe (ang. Faster R-CNN) [47] zoptymalizowały ten proces, osiągając wysoką dokładność kosztem czasu przetwarzania.

Era detektorów jednoetapowych (2016-obecnie): algorytm YOLO [26] wprowadził paradygmatyczną zmianę, traktując detekcję obiektów jako pojedynczy problem regresji. To podejście umożliwiło przetwarzanie w czasie rzeczywistym, kluczowe dla aplikacji bezpieczeństwa.

Specyficzne wyzwania detekcji w planach architektonicznych

Detekcja elementów na planach 2D różni się znacząco od standardowej detekcji obiektów w obrazach naturalnych [48, 49]:

Skala i proporcje: Różnice w skalach planów mogą wynosić kilka rzędów wielkości, od planów sytuacyjnych (skala 1:1000) po szczegóły konstrukcyjne (skala 1:10). Ta zmienność wymaga specjalnych strategii treningu i augmentacji.

Standaryzacja symboli: Brak uniwersalnych standardów oznaczenia elementów architektonicznych prowadzi do znacznej wariancji wizualnej podobnych funkcjonalnie elementów. Różni projektanci i firmy stosują odmienne konwencje graficzne.

Nakładanie się obiektów: Częste przypadki nakładających się elementów (np. drzwi w ścianach, okna w elewacjach) wymagają precyzyjnej segmentacji i post-processingu wykraczającego poza standardowe techniki NMS.

Jakość skanów: Różnorodność jakości dokumentacji źródłowej — od wysokiej rozdzielczości plików CAD po skany dokumentów archiwalnych o niskiej jakości — wymaga odporności na degradację obrazu.

3.2. Architektura YOLO — ewolucja do YOLOv8

Fundamenty architektury YOLO

Filozofia YOLO ("You Only Look Once") opiera się na traktowaniu detekcji obiektów jako pojedynczego problemu regresji [26]. W przeciwieństwie do metod dwufazowych, YOLO przewiduje jednocześnie współrzędne bounding-box i prawdopodobieństwa klas dla wszystkich obiektów w obrazie podczas pojedynczego przejścia przez sieć. Podstawowa formuła YOLO dzieli obraz wejściowy na siatkę $S \times S$ komórek, gdzie każda komórka przewiduje B bounding-box wraz z ich "confidence":

$$confidence = P(\text{Object}) \times \text{IoU}_{\text{pred}}^{\text{truth}}$$

gdzie $P(\text{Object})$ to prawdopodobieństwo obecności obiektu w komórce, a $\text{IoU}_{\text{pred}}^{\text{truth}}$ to IoU między przewidywanym a rzeczywistym bounding-box.

Ewolucja do YOLOv8 — kluczowe innowacje

YOLOv8 [50] wprowadza kilka fundamentalnych usprawnień w stosunku do poprzednich wersji:

Architektura szkieletowa

YOLOv8 wykorzystuje zmodyfikowaną architekturę CSPNet z blokami C2f (Cross Stage Partial with 2 convolutions), które oferują lepszą równowagę między dokładnością a efektywnością obliczeniową:

$$\text{C2f}(x) = \text{Concat}[\text{Conv}(x), \text{Bottleneck}^n(\text{Split}(x))]$$

Moduły C2f umożliwiają bogatszą propagację gradientów przy jednoczesnym utrzymaniu efektywności obliczeniowej.

Spatial Pyramid Pooling Fast (SPPF)

SPPF zastąpił tradycyjny SPP, oferując podobną funkcjonalność przy znacznie zredukowanym czasie obliczeń:

$$\text{SPPF}(x) = \text{Concat}[\text{MaxPool}_5^3(\text{MaxPool}_5^2(\text{MaxPool}_5(x)))]$$

Ta architektura umożliwia efektywne przetwarzanie cech w różnych skalach, kluczowe dla detekcji elementów architektonicznych o zmiennych rozmiarach.

Decoupled Head

YOLOv8 implementuje tzw. decoupled-head, gdzie klasyfikacja i lokalizacja są obsługiwane przez oddzielne ścieżki:

$$\text{Classification} = \text{Conv}_{cls}(\text{Features})$$
$$\text{Regression} = \text{Conv}_{reg}(\text{Features})$$

To rozdzielenie pozwala na lepszą optymalizację każdego zadania i prowadzi do wyższej ogólnej dokładności.

Anchor-free design

YOLOv8 rezygnuje z predefiniowanych "anchor boxes", implementując tzw. anchor-free predykcję współrzędnych:

$$x = (s_x + \sigma(t_x)) \times \text{stride}$$
$$y = (s_y + \sigma(t_y)) \times \text{stride}$$
$$w = \exp(t_w) \times \text{stride}$$
$$h = (t_h) \times \text{stride}$$

gdzie (s_x, s_y) to współrzędne komórki siatki, σ to funkcja sigmoidalna, a (t_x, t_y, t_w, t_h) to przewidywane parametry.

Optymalizacje dla planów architektonicznych

Standardowa konfiguracja YOLOv8 została dostosowana do specyfiki planów architektonicznych:

Multi-scale training: Implementacja treningu z wykorzystaniem różnych rozdzielczości wejściowych (640×640 do 1280×1280) dla lepszej generalizacji na plany o różnych skalach.

Aspect ratio augmentation: Specjalne augmentacje uwzględniające typowe proporcje planów architektonicznych (panoramiczne układy, plany kwadratowe).

Small object optimization: Modyfikacje w funkcji straty uwzględniające wysoką częstość małych obiektów (symbole, detale architektoniczne) charakterystyczną dla planów technicznych.

3.3. Konfiguracja eksperymentalna

Charakterystyka zbioru danych

Zbiór obejmuje 304 obrazy treningowe i 148 walidacyjnych, z adnotacjami dla klas: *SingleDoor*, *DLDoor*, *Window*, *IntWall*, *ExtWall*, *Stairs*, *SlidingDoor*.

Źródła:

- rzeczywiste plany architektoniczne (publiczne repozytoria, dokumentacje),
- układy syntetyczne (generacja proceduralna),
- konwersje BIM do obrazów w ujednoliconej rozdzielczości.

Adnotacja: CVAT, precyzyjne bounding-box i etykiety klas, kontrola jakości (wielu anotatorów na podzbiorze) z uzyskaną zgodnością $\kappa = 0.94$.

Konfiguracja sprzętowa i programowa

Sprzęt: RTX 3080 (10 GB VRAM), Intel i7-10700K, 32 GB RAM, NVMe 1 TB.

Oprogramowanie: Ultralytics YOLOv8 (8.3.96), Python 3.11.2 + CUDA 12.1, PyTorch 2.1.2 [51], opencv 4.8.0, NumPy 1.24.3, Pandas 2.0.3.

Uzasadnienie wyboru hiperparametrów

Wybór modelu YOLOv8m

YOLOv8m został wybrany jako kompromis między złożonością obliczeniową a jakością detekcji na podstawie wstępnych eksperymentów porównawczych:

YOLOv8m oferuje najlepszą równowagę między dokładnością (mAP@0.5 = 0.968) a wydajnością (74 fps), co jest kluczowe dla praktycznych zastosowań w analizie bezpieczeństwa.

Rozmiar wejścia: 832×832 pikseli

Wybór rozdzielczości 832×832 wynika z kompromisu między precyzją detekcji a efektywnością obliczeniową:

$$\text{Receptive Field} = \sum_{i=1}^{n} k_i \prod_{j=1}^{i-1} s_j$$

Wariant	Parametry (M)	FLOPs (G)	mAP@0.5	FPS
YOLOv8n	3.2	8.7	0.892	124
YOLOv8s	11.2	28.6	0.941	89
YOLOv8m	25.9	78.9	0.968	74
YOLOv8l	43.7	165.2	0.975	56
YOLOv8x	68.2	257.8	0.978	43

Tabela 3.1: Porównanie wariantów YOLOv8 dla detekcji elementów architektonicznych

gdzie k_i to rozmiar kernela warstwy i, a s_j to krok przesunięcia filtra (stride) warstwy j.

Dla rozdzielczości 832×832, efektywne "receptive field"wynosi około 627×627 pikseli, co pozwala na detekcję obiektów o rozmiarach od 8×8 do 400×400 pikseli — optymalny zakres dla elementów architektonicznych.

Optymalizator AdamW

Proces trenowania sieci neuronowej wymaga algorytmu optymalizacji, który kontroluje sposób aktualizacji wag modelu na podstawie obliczonych gradientów funkcji straty. Optymalizator stanowi kluczowy element procesu uczenia - określa kierunek i wielkość zmian parametrów sieci w celu minimalizacji błędów predykcji. W kontekście detekcji obiektów na planach architektonicznych, gdzie model musi nauczyć się rozpoznawać różnorodne elementy (drzwi, okna, ściany), wybór odpowiedniego algorytmu optymalizacji ma bezpośredni wpływ na jakość i stabilność treningu.

AdamW [52] został wybrany ze względu na jego przewagę nad standardowym Adamem w kontekście regulacji wagowej:

$$\theta_{t+1} = \theta_t - \alpha \left(\frac{m_t}{\sqrt{v_t} + \epsilon} + \lambda \theta_t \right)$$

gdzie: - $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$ (momentum pierwszego rzędu) - $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$ (momentum drugiego rzędu) - $\lambda = 1 \times 10^{-4}$ (współczynnik weight decay) AdamW rozdziela regulację L2 od adaptacyjnego skalowania gradientów, podczas gdy Adam dodaje regulację bezpośrednio do gradientu, co prowadzi do niepożądanych interakcji między mechanizmami adaptacyjnego uczenia a kontrolą wag. To rozdzielenie w AdamW zapewnia lepszą generalizację modelu, szczególnie istotną przy ograniczonym zbiorze danych architektonicznych.

Learning rate: 9.09×10^{-4}

Współczynnik uczenia został określony za pomocą metody *learning rate finder* [53]:

$$\mathrm{lr} = \mathrm{lr}_{min} \times \left(\frac{\mathrm{lr}_{max}}{\mathrm{lr}_{min}} \right)^{\frac{\text{epoch}}{\text{total_epochs}}}$$

Eksperymentalne określenie optymalnego zakresu wykazało, że α = 9.09×10^{-4} zapewnia stabilną konwergencję bez oscylacji charakterystycznych dla wyższych wartości.

Batch size: 16

Rozmiar batch został wybrany na podstawie ograniczeń pamięci GPU (10GB RTX 3080) oraz stabilności treningu:

$$\text{Memory Usage} = \text{batch_size} \times \text{input_size}^2 \times \text{channels} \times 4 \text{ bytes}$$

Dla konfiguracji 16×832²×3×4 ≈ 265MB na dane wejściowe, pozostawiając wystarczająco pamięci na gradienty i stany optymalizatora.

3.4. Strategia augmentacji danych

Augmentacja geometryczna

Ze względu na specyfikę planów architektonicznych, tradycyjne augmentacje zostały dostosowane:

Rotacja: Ograniczona do ±10° (w przeciwieństwie do standardowych ±45°) ze względu na konwencjonalne orientacje planów architektonicznych.

Skalowanie: Zakres 0.8–1.2× uwzględniający typową wariancję skal w dokumentacji architektonicznej.

Translacje: Maksymalnie 10% rozmiaru obrazu, zachowujące integralność układów przestrzennych.

Augmentacja specyficzna dla planów

Artefakty skanu: Symulacja szumu skanowania, degradacji JPEG i zniekształceń optycznych typowych dla digitalizacji planów archiwalnych.

Wariacje DPI: augmentacja rozdzielczości symulująca różne jakości źródłowe dokumentów (75-600 DPI).

mozaika: Łączenie fragmentów różnych planów w celu zwiększenia różnorodności kontekstualnej bez naruszania logiki architektonicznej.

3.5. Post–processing i rafinacja detekcji

Eliminacja duplikatów (IoSA)

Duplikaty, częste w gęstych układach architektonicznych, eliminujemy przez *IoSA*, lepsze od IoU przy znacznych dysproporcjach rozmiarów:

$$\text{IoSA}(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}$$

IoSA jest szczególnie efektywne dla elementów architektonicznych o różnych rozmiarach (np. drzwi vs. okna), gdzie standardowy IoU może prowadzić do false negatives.

Próg $\tau_{\text{IoSA}} = 0.50$ został określony empirycznie na podstawie analizy 200 przypadków nakładających się detekcji:

Próg IoSA	Precision	Recall
0.30	0.923	0.987
0.40	0.941	0.981
0.50	0.962	0.975
0.60	0.978	0.952
0.70	0.985	0.934

Tabela 3.2: Optymalizacja progu IoSA

Walidacja geometryczna

Implementacja reguł geometrycznych specyficznych dla planów architektonicznych:

Walidacja równoległości ścian:

$$\cos(\theta_{wall1}, \theta_{wall2}) > 0.95 \quad \text{lub} \quad \cos(\theta_{wall1}, \theta_{wall2}) < -0.95$$

Walidacja pozycji drzwi w ścianach:

$$\text{IoU}(\text{door_bbox}, \text{wall_bbox}) > 0.7$$

Konsystencja skal:

$$\left| \frac{\text{area}_{detected}}{\text{area}_{expected}} - 1 \right| < 0.3$$

Segmentacja barier

Dla obiektów nakładających się, segmentacja w punktach przecięcia wykorzystuje "intersection-midpoint":

$$M_x = \frac{X_{min}^{(int)} + X_{max}^{(int)}}{2}, \quad M_y = \frac{Y_{min}^{(int)} + Y_{max}^{(int)}}{2}$$

Ten algorytm zapewnia logiczną segmentację elementów zachowując spójność strukturalną dla dalszej analizy grafowej.

3.6. Wyniki eksperymentalne i analiza

Analiza wydajności treningu

Ogólne metryki: W inferencji przyjęto próg zaufania $conf = 0.3$ na podstawie analizy krzywych precyzja/pewność. Tabela 3.3 podsumowuje wyniki walidacyjne po 100 epokach. Osiągnięto mAP@0.50–0.95 = 0.917, precyzja = 0.968, 'recall' = 0.980.

Elementy krytyczne dla bezpieczeństwa (różne typy drzwi) uzyskały niemal idealną detekcję. Nieco niższe wyniki dla ścian (*IntWall*/*ExtWall*) wynikają z ich podobieństwa w rzutach 2D.

Klasa	Precision	Recall	mAP@0.50	mAP@0.50-0.95
SingleDoor	0.995	1.000	0.995	0.950
DLDoor	0.992	1.000	0.995	0.979
Window	0.979	0.975	0.995	0.910
IntWall	0.921	0.970	0.958	0.786
ExtWall	0.923	0.914	0.981	0.859
Stairs	0.981	1.000	0.995	0.995
SlidingDoor	0.988	1.000	0.995	0.937
Średnia	**0.968**	**0.980**	**0.988**	**0.917**

Tabela 3.3: Szczegółowe metryki wydajności po 100 epokach

Analiza konwergencji

Krzywe strat wykazują stabilną konwergencję bez oznak tzw. overfitting'u:

Bbox Loss: Spadek z 0.051 (epoka 1) do 0.008 (epoka 100) **Cls Loss:** Redukcja z 0.891 (epoka 1) do 0.142 (epoka 100) **DFL Loss:** Stabilizacja na poziomie 1.089 od epoki 85

Równoległe trajektorie loss treningowego i walidacyjnego potwierdzają właściwą regulację modelu.

Analiza macierzy pomyłek

Szczegółowa analiza macierzy pomyłek (patrz rysunek 3.1) ujawnia charakterystyczne wzorce błędów:

Najczęstsze pomyłki:

- *IntWall* ↔ *ExtWall*: 8.2% przypadków wzajemnej pomyłki wynikającej z podobieństwa wizualnego w rzucie 2D
- *SingleDoor* ↔ *SlidingDoor*: 3.1% pomyłek w przypadkach niejednoznacznych symboli drzwiowych
- *Window* — problemy z małymi oknami: 5.7% "False Negative"dla okien o powierzchni <50 px²

Studia porównawcze

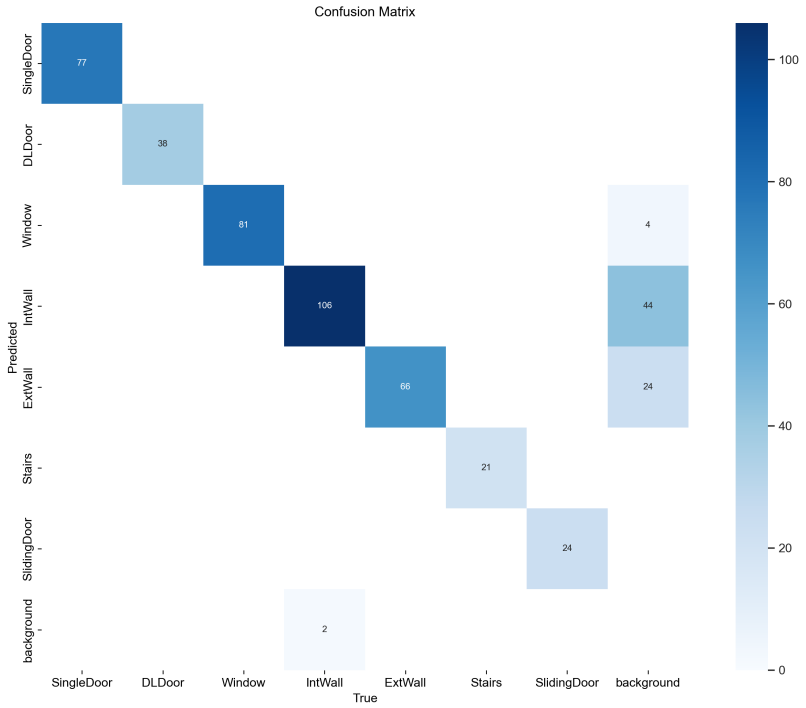
Wpływ strategii treningu

Wpływ post–processingu

3.7. Detekcja elementów systemów bezpieczeństwa

Wyzwanie niespójnych symboli na schematach

Brak standardu zapisu symboli (różni projektanci, różne konwencje) czyni klasyczne „statyczne"mapowanie symbol→znaczenie zawodnym. Ten problem rozwiązujemy poprzez autorską metodę



Rysunek 3.1: Macierz pomyłek dla modelu YOLOv8m po 100 epokach treningu. Wartości na przekątnej głównej (ciemnoniebieski kolor) reprezentują poprawne klasyfikacje dla każdej klasy, podczas gdy wartości poza przekątną wskazują na błędne przypisania. Najczęstsze pomyłki występują między klasami *IntWall* i *ExtWall* (44 i 24 przypadki wzajemnych pomyłek), co wynika z podobieństwa wizualnego tych elementów w rzucie 2D. Wysoka koncentracja wartości na przekątnej potwierdza wysoką precyzję klasyfikacji, szczególnie dla elementów krytycznych dla bezpieczeństwa jak różne typy drzwi (*SingleDoor*: 77/77, *DLDoor*: 38/38, *SlidingDoor*: 24/24 poprawnych klasyfikacji). Intensywność koloru odpowiada liczbie przypadków, gdzie ciemniejsze odcienie wskazują na większą liczbę detekcji.

Konfiguracja	mAP@0.50	mAP@0.50-0.95
Baseline (bez augmentacji)	0.932	0.823
+ mozaika	0.961	0.871
+ MixUp	0.974	0.889
+ Transformacje geometryczne	0.982	0.904
+ augmentacja architektoniczna	0.986	0.912
+ Pełna konfiguracja	**0.988**	**0.917**

Tabela 3.4: Porównanie wariantów strategii treningowych

Komponent	Precision	False Positive Rate
Surowe wyjście YOLO	0.891	0.089
+ Usuwanie duplikatów (IoU)	0.923	0.067
+ Usuwanie duplikatów (IoSA)	0.941	0.051
+ walidacja-geometryczna	0.962	0.032
+ Segmentacja barier	**0.968**	**0.024**

Tabela 3.5: Porównanie komponentów post–processingu

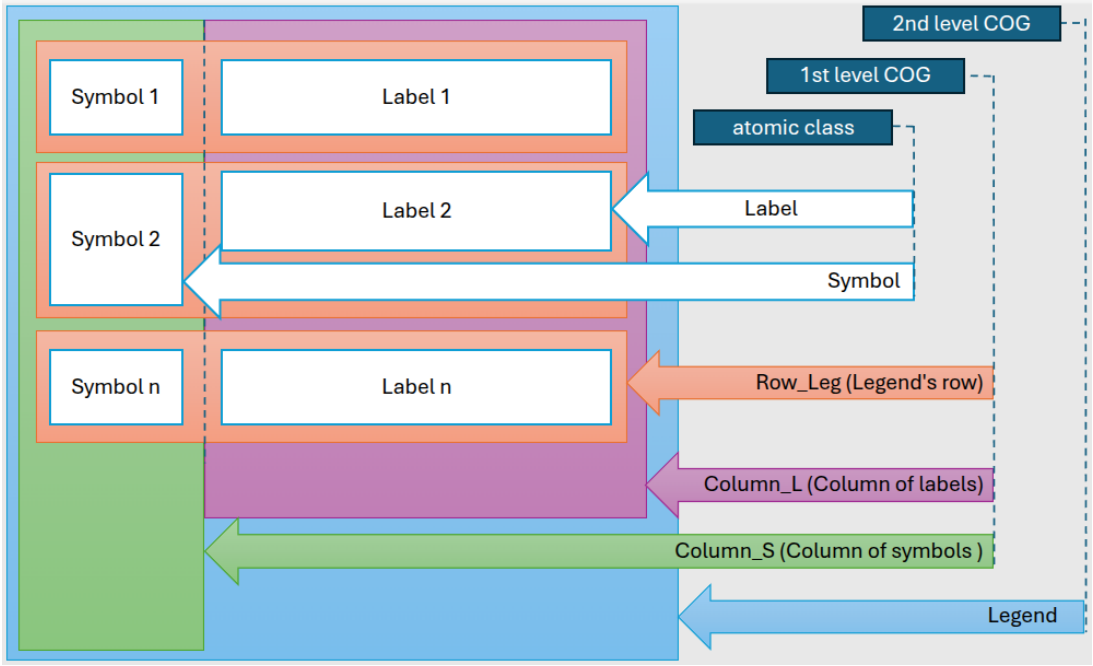
Contextual Object Grouping *COG*, która umożliwia uczenie się interpretacji symboli *w kontekście użytej legendy*.

Proces COG — interpretacja przez kontekst

COG traktuje pary *symbol–etykieta* w legendzie jako obiekty kontekstowe ("row_leg") wykrywane wprost, a następnie przenosi ich semantykę na plan [27]. Podejście to fundamentalnie różni się od tradycyjnych metod detekcji obiektów, które najpierw wykrywają poszczególne elementy atomowe, a następnie próbują wywnioskować ich wzajemne relacje za pomocą algorytmów post-processingu.

Struktura przedstawiona na rysunku 3.2 ilustruje kluczową innowację procesu COG: wprowadzenie pośredniego poziomu ontologicznego między detekcją obiektów atomowych a interpretacją semantyczną wysokiego poziomu. Klasy COG pierwszego poziomu (Row_Leg) enkapsulują strukturalne relacje przestrzenne i funkcjonalne, umożliwiając modelowi bezpośrednie postrzeganie kontekstu zamiast jego wnioskowania przez kolejne etapy przetwarzania.

Tego typu podejście pozwala na dynamiczną interpretację symboli niezależnie od standardów rysunkowych zastosowanych przez różnych projektantów. Model uczy się rozpoznawać pary symbol-etykieta jako całości semantyczne, co eliminuje potrzebę utrzymywania stałych słowników symboli i umożliwia adaptację do różnorodnych konwencji graficznych występujących w dokumentacji technicznej systemów bezpieczeństwa.



Rysunek 3.2: Hierarchiczna struktura procesu Contextual Object Grouping (COG) ilustrująca wielopoziomową organizację semantyczną w interpretacji diagramów technicznych. Diagram pokazuje, jak klasy atomowe (Symbol i Label, oznaczone kolorem jasnoniebieskim) są łączone w klasy COG pierwszego poziomu Row_Leg (pomarańczowe), reprezentujące zunifikowane pary symbol-etykieta w wierszach legendy. Te z kolei grupowane są w struktury wyższego poziomu: Column_S (zielone) agregujące wszystkie symbole, Column_L (fioletowe) agregujące wszystkie etykiety, oraz Legend (jasnozielone) reprezentujące kompletną strukturę kontekstową. Zagnieżdżone prostokąty z różnymi poziomami przezroczystości wizualizują relacje zawierania, podczas gdy strzałki wskazują hierarchiczne zależności między klasami. Źródło: opracowanie własne na podstawie [27].

Architektura

Jeden model YOLOv8m uczy się klas atomowych ("symbol", "label") oraz klas kontekstowych ("row_leg", "legend"). Dwa etapy pracy:

- **Legenda:** detekcja i strukturalizacja, ekstrakcja parowań symbol–etykieta, budowa hierarchii (JSON).
- **Plan:** wyszukiwanie symboli zgodnie z legendą i przypisywanie im znaczeń semantycznych.

Wyniki COG

- mAP@0.50 ≈ 0.99 (we wszystkich klasach),
- mAP@0.50-0.95 ≈ 0.81,
- 100% poprawnych parowań symbol–etykieta w testach,
- średnio: precyzja 96.8%, recall 98.0%.

Kluczowe obserwacje: model wykazuje *kontekstową świadomość* — wykrywa symbole w obrębie legendy, lecz nie w izolacji na planie (zgodnie z założeniem „najpierw kontekst, potem interpretacja").

3.8. Rozważania implementacyjne

Optymalizacja wydajnościowa

System można dostosować do różnorodnych wymagań praktycznych poprzez odpowiednią konfigurację parametrów przetwarzania. W trybie czasu rzeczywistego redukcja rozmiaru wejścia do 640×640 pikseli umożliwia przetwarzanie przekraczające 30 klatek na sekundę, przy niewielkim spadku średniej precyzji detekcji około 3%. Dla zastosowań wymagających najwyższej dokładności, zwiększenie rozdzielczości do 1024×1024 pikseli poprawia detekcję drobnych elementów architektonicznych o około 5%, kosztem wydłużenia czasu przetwarzania.

Przetwarzanie wsadowe planów architektonicznych wykazuje optymalne wykorzystanie zasobów GPU przy rozmiarze partii równym 8 dla dokumentów standardowych rozmiarów. Taka konfiguracja zapewnia równowagę między przepustowością systemu, a efektywnym wykorzystaniem pamięci graficznej, co jest szczególnie istotne podczas analizy dużych portfeli obiektów.

Integracja z istniejącymi systemami

Architektura systemu została zaprojektowana z uwzględnieniem kompatybilności z szeroko stosowanymi rozwiązaniami branżowymi. Bezpośrednie przetwarzanie standardowych formatów plików architektonicznych, takich jak DWG, DXF oraz PDF, eliminuje konieczność konwersji danych wejściowych i upraszcza proces wdrożenia.

Interfejsy programistyczne oparte na architekturze RESTful API umożliwiają integrację z istniejącymi systemami zarządzania bezpieczeństwem oraz platformami analitycznymi. System wspiera również łączność z bazami danych 'Building Information Modeling' (BIM), co pozwala na wzbogacenie analizy o dodatkowe informacje o właściwościach materiałowych i funkcjonalnych elementów budynku.

Obecna implementacja wykazuje znaczące różnice w skuteczności detekcji w zależności od orientacji wykrywanych symboli. W przypadku planu architektonicznego (rysunek 3.3), gdzie symbole zachowują orientację zgodną z legendą, system osiąga praktycznie perfekcyjne wyniki detekcji z wartościami ufności przekraczającymi 0.82 dla wszystkich elementów funkcjonalnych [27].

Natomiast analiza planu bezpieczeństwa (rysunek 3.4) ujawnia wrażliwość systemu na rotację symboli elementów ESS. Czujniki PIR (zielone strzałki) oraz detektory tłuczenia szkła GBD (niebieskie prostokąty) wykazują zróżnicowaną skuteczność wykrywania w zależności od ich orientacji względem położenia w legendzie. Ten aspekt stanowi kluczowy obszar wymagający dalszego rozwoju algorytmów w celu zwiększenia niezmienności względem transformacji geometrycznych.

Jak ilustruje rysunek 3.3, w przypadku planu architektonicznego, gdzie symbole zachowują orientację zgodną z legendą, system osiąga praktycznie perfekcyjne wyniki detekcji z wartościami ufności przekraczającymi 0.82 dla wszystkich elementów funkcjonalnych [27].

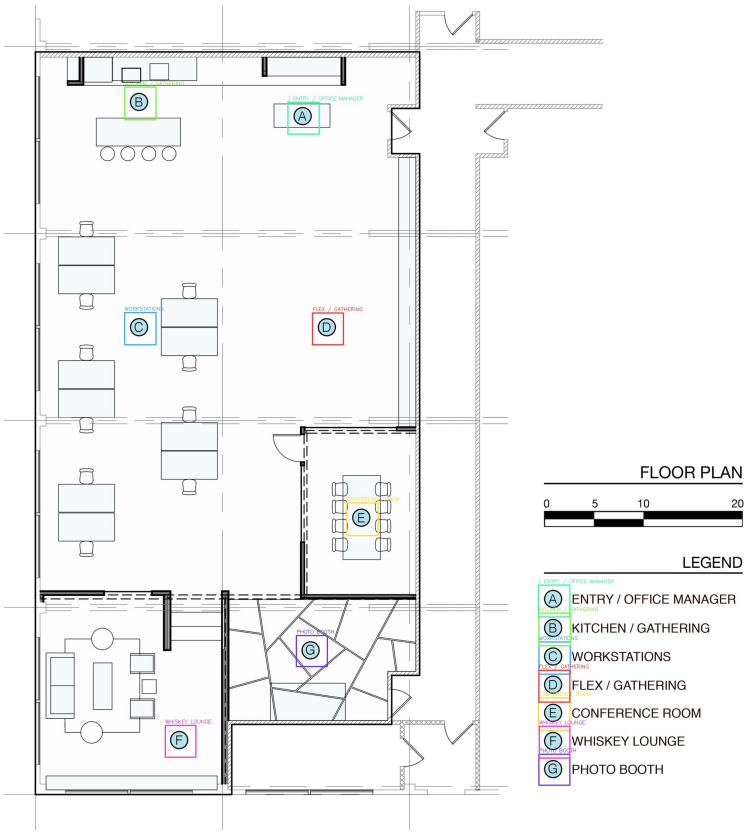
Natomiast analiza planu bezpieczeństwa (rysunek 3.4) ujawnia wrażliwość systemu na rotację symboli elementów ESS. Czujniki PIR (zielone strzałki) oraz detektory tłuczenia szkła GBD (niebieskie prostokąty) wykazują zróżnicowaną skuteczność wykrywania w zależności od ich orientacji względem położenia w legendzie.

Ten aspekt stanowi kluczowy obszar wymagający dalszego rozwoju algorytmów w celu zwiększenia niezmienności względem transformacji geometrycznych. Wpływ orientacji symboli na skuteczność detekcji elementów architektonicznych. (a) Plan architektoniczny z symbolami zachowującymi orientację zgodną z legendą - model osiąga wysokie wyniki ufności (0.82-0.99) dla wszystkich elementów oznaczonych literami A-G. (b) Plan bezpieczeństwa z elementami ESS w różnych orientacjach - zielone strzałki (czujniki PIR) i niebieskie prostokąty (detektory GBD) wykazują zróżnicowaną skuteczność detekcji w zależności od stopnia rotacji względem orientacji bazowej z legendy. Analiza pokazuje konieczność dalszego rozwoju algorytmów w kierunku większej niezmienności względem transformacji geometrycznych. Źródło: opracowanie własne na podstawie [27].

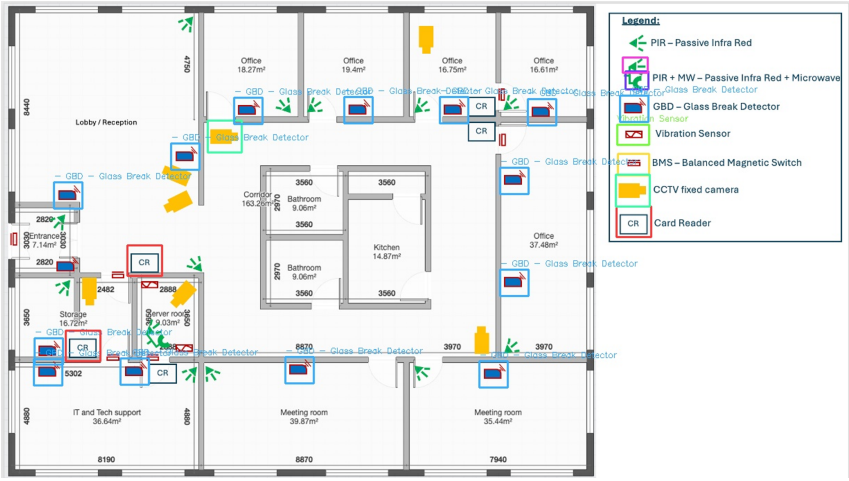
3.9. Kierunki dalszego rozwoju

Ulepszenia algorytmiczne

Rozwój systemu detekcji elementów architektonicznych może obejmować integrację z bardziej zaawansowanymi wariantami architektury YOLO, takimi jak YOLOv8x, co potencjalnie zwiększy precyzję detekcji w scenariuszach o wysokich wymaganiach dokładnościowych. Implementacja dynamicznego doboru progów ufności w zależności od typu wykrywanego elementu mogłaby dodatkowo poprawić równowagę między czułością a specyficznością systemu.



Rysunek 3.3: Precyzyjne mapowanie symboli zachowujących orientację z legendy. Plan architektoniczny z elementami oznaczonymi literami A-G, gdzie model osiąga wysokie wyniki ufności (0.82-0.99) dla wszystkich wykrywanych symboli dzięki zachowaniu spójności orientacyjnej między legendą a głównym diagramem. Źródło: opracowanie własne na podstawie [27].



Rysunek 3.4: Wrażliwość na rotację symboli bezpieczeństwa. Plan z elementami ESS w różnych orientacjach, gdzie zielone strzałki (czujniki PIR) i niebieskie prostokąty (detektory GBD) wykazują zróżnicowaną skuteczność detekcji w zależności od stopnia rotacji względem orientacji bazowej z legendy. Analiza pokazuje konieczność dalszego rozwoju algorytmów w kierunku większej niezmienności względem transformacji geometrycznych. Źródło: opracowanie własne na podstawie [27].

Istotnym kierunkiem rozwoju jest rozszerzenie reprezentatywności bazy danych treningowych o plany architektoniczne z różnych regionów geograficznych, co zwiększy zdolność generalizacji modelu na różnorodne standardy projektowe i konwencje rysunkowe stosowane na świecie. Szczególnie ważne jest uwzględnienie różnorodnych stylów architektonicznych i tradycji projektowych charakterystycznych dla różnych kultur i systemów prawnych.

Rozszerzenia funkcjonalne

Integracja z modelami trójwymiarowymi, w tym chmurami punktów i systemami BIM, stanowi naturalną ścieżkę ewolucji systemu w kierunku pełnej analizy przestrzennej obiektów. Automatyczne rozpoznawanie skali planu architektonicznego mogłoby wyeliminować konieczność manualnej kalibracji systemu dla dokumentów o różnych skalach przedstawienia.

Implementacja mechanizmów uczenia transferowego dla specyficznych typów budynków (przemysłowe, mieszkaniowe, użyteczności publicznej) pozwoliłaby na specjalizację modelu dla różnych domen architektonicznych. Rozszerzenie procesu COG o hierarchiczne struktury interpretacji symboli mogłoby zwiększyć elastyczność systemu w zakresie adaptacji do nowych standardów dokumentacji technicznej.

Rozwój algorytmów niezmienności względem transformacji geometrycznych stanowi priorytetowy kierunek badań z uwagi na wspomnianą wrażliwość metody na rotację symboli, co pokazano na rysunku 3.4. Implementacja technik augmentacji danych obejmujących rotacje, skalowanie i transformacje perspektywiczne mogłaby znacząco poprawić odporność systemu na różnorodne orientacje symboli występujące w rzeczywistych planach architektonicznych.

Perspektywy długoterminowe

Przyszły rozwój systemu może obejmować integrację z systemami Cyber-Physical Protection Systems w czasie rzeczywistym, umożliwiając dynamiczną aktualizację modeli bezpieczeństwa w odpowiedzi na zmiany w konfiguracji fizycznej obiektu. Automatyczna adaptacja do ewoluujących standardów architektonicznych poprzez mechanizmy samo-uczenia mogłaby zapewnić długoterminową aktualność systemu.

Wykorzystanie tzw. foundation models do interpretacji planów architektonicznych w języku naturalnym stanowi perspektywiczny kierunek rozwoju, który mógłby umożliwić bardziej intuicyjne interfejsy użytkownika oraz automatyczne generowanie opisów tekstowych analizowanych struktur. Takie rozwiązania mogłyby znacząco rozszerzyć dostępność systemu dla specjalistów spoza dziedziny technologii komputerowych.

Integracja z systemami rzeczywistości rozszerzonej (AR) i wirtualnej (VR) mogłaby umożliwić immersyjną wizualizację wyników analizy bezpieczeństwa, ułatwiając zrozumienie przestrzennych relacji między elementami systemu ochrony. Połączenie automatycznej detekcji z interaktywnymi środowiskami wizualizacji mogłoby znacząco zmienić sposób projektowania i oceny systemów bezpieczeństwa obiektów.

3.10. Podsumowanie

System detekcji elementów architektonicznych na planach 2D osiąga jakość operacyjną (mAP@0.50-0.95 = 0.917) przy niskim FPR po post-processingu. Zaawansowane techniki (IoSA, walidacja geometryczna, segmentacja barier) wyraźnie poprawiają precyzję (0.968) i redukują FPR (0.024).

Kluczowe osiągnięcia obejmują:

- Wysoką dokładność detekcji elementów krytycznych dla bezpieczeństwa (>99% dla drzwi)
- Efektywną interpretację symboli systemów bezpieczeństwa poprzez COG
- Skalowalną architekturę umożliwiającą przetwarzanie dużych zbiorów planów
- Solidne podstawy dla automatycznej analizy ścieżek ataku przedstawionej w kolejnych rozdziałach

Warstwa detekcji dostarcza dwie komplementarne funkcje: (1) wiarygodne wykrywanie elementów architektonicznych z planów 2D oraz (2) automatyczną interpretację symboli infrastruktury bezpieczeństwa w oparciu o kontekst legendy. Tworzy to kompletną, zautomatyzowaną warstwę pozyskiwania danych wejściowych do dalszych etapów analizy przedstawionych w kolejnych rozdziałach.

4. Konstruowanie Grafów Obiektów i Heurystyczna Ocena Ścieżek Ataku z Wykorzystaniem Algorytmu A*

Rozdział poprzedni koncentrował się na warstwie percepcyjnej, czyli na automatycznej detekcji elementów infrastruktury budynków w planach technicznych z wykorzystaniem algorytmów komputerowej analizy obrazu. Dzięki temu uzyskano możliwość systematycznego i zautomatyzowanego pozyskiwania danych wejściowych o barierach, punktach dostępu i innych obiektach istotnych z punktu widzenia bezpieczeństwa. Jednak sama identyfikacja elementów nie wystarcza do oceny podatności obiektu. Konieczne jest również ich powiązanie w spójną strukturę analityczną, która pozwoli modelować potencjalne działania przeciwnika. Dlatego w niniejszym rozdziale omawiane są metody generowania grafów ścieżek ataku oraz ich heurystycznej analizy z wykorzystaniem algorytmów przeszukiwania i technik topologicznych, stanowiących pomost między surowymi danymi detekcyjnymi a oceną skuteczności systemu ochrony.

4.1. Wprowadzenie teoretyczne do teorii grafów

Podstawy teorii grafów w kontekście bezpieczeństwa

Teoria grafów stanowi fundamentalną dziedzinę matematyki dyskretnej, która znajduje szerokie zastosowanie w modelowaniu systemów bezpieczeństwa fizycznego. Graf $G = (V, E)$ składa się z niepustego zbioru wierzchołków V i zbioru krawędzi E, gdzie każda krawędź łączy parę wierzchołków [54, 55]. W kontekście analizy bezpieczeństwa obiektów grafy umożliwiają reprezentację przestrzeni architektonicznej jako sieci połączonych elementów, gdzie wierzchołki reprezentują lokalizacje przestrzenne lub elementy infrastruktury, a krawędzie modelują możliwe przejścia lub relacje między nimi [56, 57].

Proces tworzenia grafów bezpieczeństwa rozpoczyna się od analizy rzeczywistego układu fizycznego obiektu. Rysunek 4.1 przedstawia przykładową strukturę obiektu z nałożonymi węzłami reprezentującymi kluczowe elementy systemu ochrony fizycznej.

Praktyczna implementacja tej koncepcji ilustruje rysunek 4.2, przedstawiający transformację fizycznego układu obiektu w abstrakcyjny graf zawierający wszystkie względnie racjonalne ścieżki ataku.

Matematycznie, graf można zapisać jako $G = (V, E)$, gdzie $V = \{v_1, v_2, \ldots, v_n\}$ stanowi zbiór wierzchołków, a $E \subseteq V \times V$ reprezentuje zbiór krawędzi. Jak pokazano na rysunkach 4.1 i 4.2, w aplikacjach bezpieczeństwa funkcja wagi $w : E \to \mathbb{R}^+$ przypisuje każdej krawędzi wartość reprezentującą koszt, czas, ryzyko lub inne metryki istotne dla analizy zagrożeń [58]. Transformacja z układu fizycznego

Rysunek 4.1: Układ fizyczny obiektu z nałożonymi węzłami systemu bezpieczeństwa. Struktura przedstawia hipotetyczny obiekt chroniony z zaznaczonymi elementami: punkty perymetryczne (F1-F11) na granicy obiektu, elementy ścian zewnętrznych i wewnętrznych (W51-W59), punkty dostępu (D31-D43), oraz cel ataku (T). Każdy węzeł reprezentuje strategiczne miejsce w systemie ochrony, gdzie może dojść do detekcji, opóźnienia lub neutralizacji działań przeciwnika.

do modelu grafowego umożliwia zastosowanie zaawansowanych algorytmów analizy sieci do identyfikacji podatności systemów ochrony oraz optymalizacji strategii bezpieczeństwa.

Klasyfikacja grafów w modelowaniu bezpieczeństwa

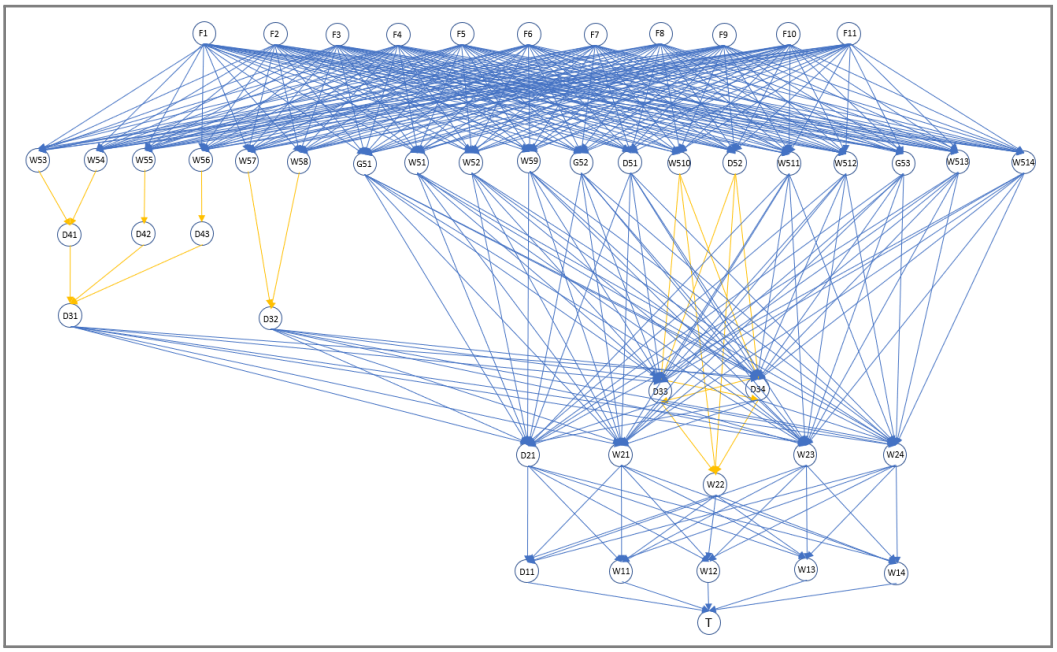
Grafy skierowane vs. nieskierowane w analizie zagrożeń

Wybór między grafami skierowanymi i nieskierowanymi ma kluczowe znaczenie dla dokładności modelowania rzeczywistych scenariuszy bezpieczeństwa. Tradycyjne podejścia w analizie ryzyka często wykorzystują grafy nieskierowane, zakładając symetrię przejść między elementami systemu [59, 60]. Jednak współczesne badania wskazują na znaczące korzyści płynące z zastosowania grafów skierowanych w modelowaniu systemów bezpieczeństwa fizycznego [61, 62].

Uzasadnienie wykorzystania grafów skierowanych:

1. **Asymetria przejść bezpieczeństwa**: Elementy architektoniczne wykazują często asymetrię w możliwościach przejścia. Jednokierunkowe drzwi awaryjne, systemy kontroli dostępu z weryfikacją biometryczną, czy bariery przeciwpożarowe naturalnie wymuszają reprezentację skierowaną [7, 63].

2. **Modelowanie barier jednokierunkowych**: Zawory bezpieczeństwa, śluzy dekontaminacyjne, czy systemy bezpieczeństwa z tzw. mechanizmami anti-tailgating wymagają precyzyjnej reprezentacji kierunkowości w grafach [64].



Rysunek 4.2: Graf bezpieczeństwa powstały na podstawie mapowania obiektu z rysunku 4.1 na sieć wszystkich względnie racjonalnych ścieżek ataku. Węzły zachowują oznaczenia z układu fizycznego: punkty perimetryczne (F1-F11), elementy ścian (W51-W59), przejścia drzwiowe (D31-D43) oraz cel ataku (T). Krawędzie ilustrują możliwe trasy przemieszczania się przeciwnika, z rozróżnieniem na standardowe połączenia (niebieskie linie) i ścieżki wymagające pokonania dodatkowych barier (żółte linie). Graf demonstruje złożoność analizy wszystkich potencjalnych scenariuszy infiltracji w systemie CPPS oraz umożliwia zastosowanie algorytmów teorii grafów do znajdowania optymalnych ścieżek ataku i identyfikacji krytycznych węzłów systemu.

3. **Kontekst operacyjny i narzędzia przeciwnika**: Przeciwnik może mieć różne możliwości pokonania przeszkody w zależności od kierunku podejścia. Pokonanie okna od strony wewnętrznej wymaga innych narzędzi i czasu niż od strony zewnętrznej [65, 66].

4. **Precyzja modelowania procedur bezpieczeństwa**: Procedury ewakuacji, protokoły dostępu, czy sekwencje blokad bezpieczeństwa wykazują inherentną kierunkowość, którą grafy nieskierowane nie mogą precyzyjnie reprezentować [67].

Formalnie, w grafie skierowanym krawędź $(u, v) \neq (v, u)$, co pozwala na niezależne modelowanie kosztów i ograniczeń dla każdego kierunku przejścia [68].

Właściwości strukturalne grafów w analizie bezpieczeństwa

Spójność i dostępność

Spójność grafu w kontekście bezpieczeństwa determinuje dostępność różnych obszarów dla potencjalnego intruza. Graf jest silnie spójny, jeśli istnieje ścieżka skierowana między każdą parą wierzchołków [69]. W analizie bezpieczeństwa rozróżniamy:

– **Spójność silna**: Wszystkie obszary są wzajemnie dostępne, co może wskazywać na braki w segmentacji bezpieczeństwa

– **Spójność słaba**: Niektóre obszary są dostępne jedynie w określonych kierunkach, co może reprezentować właściwe implementacje stref bezpieczeństwa [70]

Centralność i węzły krytyczne

Analiza centralności węzłów pozwala zidentyfikować krytyczne punkty w systemie bezpieczeństwa [71, 72]:

– **Centralność pośredniczenia (betweenness centrality)**: Identyfikuje węzły leżące na największej liczbie najkrótszych ścieżek między innymi węzłami

– **Centralność bliskości (closeness centrality)**: Wskazuje węzły o najkrótszych średnich odległościach do wszystkich innych węzłów

– **Centralność stopnia (degree centrality)**: Określa węzły o największej liczbie połączeń bezpośrednich

Cykliczność i redundancja ścieżek

Obecność cykli w grafie architektonicznym ma dwoisty charakter w kontekście bezpieczeństwa [73]:

– **Pozytywny aspekt**: Redundancja ścieżek ewakuacji, alternatywne trasy dla służb ratunkowych

– **Negatywny aspekt**: Alternatywne ścieżki obejścia zabezpieczeń, możliwości manewrowania intruza

Reprezentacja grafów - analiza efektywności struktury danych

Wybór struktury danych dla reprezentacji grafu ma istotny wpływ na efektywność algorytmów analizy bezpieczeństwa [74, 75].

Macierz sąsiedztwa oferuje stały czas dostępu $O(1)$ do informacji o krawędzi, ale wymaga $O(|V|^2)$ pamięci, co staje się prohibitywne dla dużych struktur architektonicznych [76]. **Listy sąsiedztwa** zapewniają efektywność pamięciową $O(|V|+|E|)$, szczególnie istotną dla rzadkich grafów architektonicznych, gdzie typowy węzeł łączy się z ograniczoną liczbą sąsiadów [77].

Współczesne badania wskazują na hybrydowe struktury danych, łączące zalety obu podejść dla specyficznych aplikacji analizy bezpieczeństwa [78, 79].

4.2. Algorytmy przeszukiwania grafów - ewolucja i zastosowania

Klasyczne algorytmy nieinformowane

Przeszukiwanie wszerz (BFS)

Algorytm przeszukiwania wszerz systematycznie eksploruje graf poziom po poziomie, gwarantując znalezienie najkrótszej ścieżki dla grafów o jednakowych wagach krawędzi [80]. W kontekście analizy bezpieczeństwa, Breadth-First Search (BFS) znajduje zastosowanie w:

- Identyfikacji minimalnej liczby barier do pokonania
- Analizie ścieżek ewakuacji o równomiernym koszcie przemieszczenia
- Określeniu stref bezpośredniej dostępności z punktów perimetrycznych [81]

Złożoność obliczeniowa $O(|V|+|E|)$ czyni BFS efektywnym dla analiz wymagających kompletności przeszukiwania przy ograniczonych zasobach obliczeniowych [82].

Przeszukiwanie w głąb (DFS)

Depth-First Search eksploruje graf, zagłębiając się maksymalnie wzdłuż każdej gałęzi przed wycofaniem [69]. W analizie bezpieczeństwa wykorzystywany jest do:

- Enumeracji wszystkich możliwych ścieżek między punktami
- Wykrywania cykli i analiza redundancji zabezpieczeń
- Identyfikacji składowych silnie spójnych w grafach zabezpieczeń [83]

Algorytmy ważone - Dijkstra

Algorytm Dijkstry stanowi fundamentalny algorytm znajdowania najkrótszej ścieżki w grafach z nieujemnymi wagami krawędzi [84]. Wprowadzony w 1959 roku, algorytm wykorzystuje zachłanną strategię eksploracji, gwarantując optymalność rozwiązania dla przypadków bez ujemnych wag [85].

Właściwości algorytmu Dijkstry w kontekście bezpieczeństwa:

– **Optymalność**: Gwarancja znalezienia najkrótszej ścieżki przy nieujemnych wagach krawędzi

– **Kompletność**: Eksploracja wszystkich dostępnych ścieżek do momentu znalezienia optimum

– **Złożoność**: $O((|V| + |E|) \log |V|)$ z kopcem Fibonacciego, $O(|V|^2)$ z prostą implementacją [86]

Ograniczenia w aplikacjach bezpieczeństwa:

Algorytm Dijkstry nie uwzględnia informacji heurystycznych o kierunku poszukiwań, co prowadzi do równomiernej eksploracji we wszystkich kierunkach [87]. To ograniczenie staje się szczególnie problematyczne w dużych grafach reprezentujących kompleksy architektoniczne, gdzie preferowana jest eksploracja ukierunkowana na cel [88].

4.3. Algorytm A* - teoria i zaawansowane zastosowania

Podstawy teoretyczne A*

Algorytm A* (A-star), opracowany przez Harta, Nilssona i Raphaela w 1968 roku, stanowi rozszerzenie algorytmu Dijkstry o komponent heurystyczny [89]. Stanowi klasyczny przykład algorytmu przeszukiwania informowanego, wykorzystującego funkcję heurystyczną do kierowania procesu eksploracji w stronę celu [90, 91].

Funkcja kosztu A*:

$$f(n) = g(n) + h(n)$$

gdzie:

– $g(n)$ - rzeczywisty koszt najkrótszej znanej ścieżki od węzła startowego do węzła n

– $h(n)$ - heurystyczna ocena kosztu od węzła n do celu

– $f(n)$ - oszacowany całkowity koszt najlepszej ścieżki przechodzący przez węzeł n

Właściwości funkcji heurystycznej

Admisyjność (Admissibility)

Funkcja heurystyczna $h(n)$ jest **admisyjna**, jeśli nigdy nie przecenia rzeczywistego kosztu dotarcia do celu [92, 93]:

$$\forall n : h(n) \leq h^*(n)$$

gdzie $h^*(n)$ reprezentuje rzeczywisty minimalny koszt od n do celu.

Twierdzenie o admisyjności: A* z admisyjną heurystyką zawsze znajdzie optymalną ścieżkę, jeśli taka istnieje [94]. Dowód opiera się na fakcie, że A* nigdy nie pominie optymalnej ścieżki z powodu nadmiernej estymacji kosztu przez funkcję heurystyczną.

Spójność (Consistency/Monotonicity)

Funkcja heurystyczna jest **spójna** (lub monotoniczna), jeśli spełnia warunek trójkąta [95]:

$$\forall n, m : h(n) \leq c(n, m) + h(m)$$

gdzie $c(n, m)$ reprezentuje koszt krawędzi od węzła n do węzła m.

Implikacje spójności:

- Spójność implikuje admisyjność, ale nie na odwrót [96]
- A* ze spójną heurystyką nigdy nie będzie musiał ponownie rozważać węzłów (właściwość monotonii f-wartości wzdłuż ścieżek) [97]
- Gwarantuje efektywność obliczeniową poprzez eliminację redundantnych obliczeń [98]

Zaawansowane warianty A*

Ważone A* (WA*)

Ważone A* modyfikuje standardową funkcję kosztu przez wprowadzenie współczynnika wagowego $\epsilon \geq 1$ [99]:

$$f(n) = g(n) + \epsilon \cdot h(n)$$

Parametr ϵ kontroluje balans między optymalnością, a szybkością:

- $\epsilon = 1$: Standardowy A* z gwarancją optymalności
- $\epsilon > 1$: Szybsze wykonanie kosztem potencjalnej sub-optymalności [100]

Anytime A*

Anytime A* umożliwia progresywne poprawianie jakości rozwiązania w miarę dostępności dodatkowego czasu obliczeniowego [101]. Algorytm początkowo znajduje sub-optymalne rozwiązanie, które następnie systematycznie ulepsza.

Bidirectional A*

Dwukierunkowe przeszukiwanie A* jednocześnie eksploruje przestrzeń od węzła startowego i docelowego, znacznie redukując przestrzeń przeszukiwania dla wielu aplikacji [102, 103].

4.4. Funkcje heurystyczne w analizie bezpieczeństwa

Klasyczne heurystyki geometryczne

Odległość euklidesowa

Dla przestrzeni dwuwymiarowej, heurystyka euklidesowa definiowana jest jako [104]:

$$h_{euclidean}(n, goal) = \sqrt{(x_n - x_{goal})^2 + (y_n - y_{goal})^2}$$

Zastosowania w bezpieczeństwie:

– Analiza przestrzeni otwartych bez barier architektonicznych

– Szacowanie czasu przemieszczenia w liniach prostych

– Baseline dla porównania z bardziej złożonymi metrykami [105]

Odległość Manhattan

W środowiskach architektonicznych z ograniczeniami przemieszczenia do kierunków ortogonalnych, odległość Manhattan często lepiej odzwierciedla rzeczywiste wzorce przemieszczania [106]:

$$h_{manhattan}(n, goal) = |x_n - x_{goal}| + |y_n - y_{goal}|$$

Uzasadnienie teoretyczne: Badania w zakresie analizy ruchów w budynkach wskazują, że wzorce przemieszczania są w znacznej mierze ograniczone do kierunków wyznaczonych przez korytarze, przejścia i struktury architektoniczne [107, 108]. Odległość Manhattan lepiej aproksymuje te ograniczenia niż metryka euklidesowa.

Specjalizowane heurystyki dla bezpieczeństwa

Teoretyczne podstawy własnej heurystyki ("stealth")

Opracowana dla potrzeb analizy ścieżek ataku, heurystyka "stealth"integruje czynniki wpływające na atrakcyjność węzła z perspektywy minimalizacji ryzyka wykrycia [3, 46]:

$$h_{stealth}(n, t) = w_1 \cdot D_{Manhattan}(n, t) \cdot S_{scale} + w_2 \cdot (1/DD_n) + w_3 \cdot BT_n + w_4 \cdot DR_n$$

Komponenty funkcji:

1. **Komponent odległości** ($w_1 \cdot D_{Manhattan}(n, t) \cdot S_{scale}$): Podstawowa orientacja przestrzenna z realistyczną metryką odległości i skalowaniem wymiarów

2. **Komponent opóźnienia detekcji** ($w_2 \cdot 1/DD_n$): Preferuje węzły o większym opóźnieniu poprzez odwrotność wartości, odzwierciedlając strategię unikania szybkiej detekcji

3. **Komponent czasu bariery** ($w_3 \cdot BT_n$): Penalizuje węzły wymagające długiego czasu pokonania, uwzględniając preferencje atakującego dla szybkich przejść

4. **Komponent ryzyka detekcji** ($w_4 \cdot DR_n$): Bezpośrednia penalizacja węzłów o wysokim inherentnym ryzyku wykrycia

Kalibracja parametrów wagowych

Współczynniki wagowe w_1, w_2, w_3, w_4 umożliwiają modelowanie różnych scenariuszy przeciwnika poprzez dostrojenie priorytetów [2]:

Scenariusz "stealth professional": Wysokie w_2 i w_4 odzwierciedlają priorytet unikania detekcji charakterystyczny dla doświadczonych adwersarzy z zaawansowanymi narzędziami [9].

Scenariusz "time-critical attack": Dominujące w_3 reprezentuje priorytet szybkości wykonania, typowy dla ataków pod presją czasu lub skoordynowanych operacji [10].

Scenariusz "amateur intruder": Zbalansowane wagi z preferencją w_3 odzwierciedlają ograniczone umiejętności i podstawowe narzędzia [11].

4.5. Teoria gier w modelowaniu przeciwnika

Modele teoriogrowe w analizie zagrożeń

Teoria gier w bezpieczeństwie

Modelowanie interakcji między obrońcą a atakującym można formalizować jako grę dwuosobową o sumie niezerowej [109, 110]. Macierz wypłat $U = [u_{ij}]$ reprezentuje użyteczność dla gracza przy wyborze strategii i przez obrońcę i strategii j przez atakującego.

Równowaga Nasha w kontekście bezpieczeństwa: Para strategii (s^*, t^*) stanowi równowagę Nasha, jeśli żaden z graczy nie może jednostronnie poprawić swojej wypłaty poprzez zmianę strategii [111, 112].

Stackelberg Security Games

W modelu Stackelberga obrońca (lider) podejmuje decyzje pierwsze, a atakujący (follower) reaguje optymalnie na zaobserwowane działania [113]. Model ten lepiej odzwierciedla rzeczywiste scenariusze bezpieczeństwa, gdzie systemy ochrony są instalowane przed potencjalnym atakiem.

Modelowanie niepewności i niepełnej informacji

Grafy z niepewnymi parametrami

W rzeczywistych scenariuszach bezpieczeństwa, parametry systemu (prawdopodobieństwa detekcji, czasy pokonania barier) często charakteryzują się niepewnością [114, 115]. Formalizacja poprzez rozkłady prawdopodobieństwa:

$$P(\text{detection}) \sim \text{Beta}(\alpha, \beta)$$
$$T(\text{barrier}) \sim \text{LogNormal}(\mu, \sigma^2)$$

Uwaga metodologiczna: Powyższe zagadnienia teorii gier i modelowania niepewności stanowią rozległą dziedzinę badawczą, którą w niniejszej pracy traktujemy jedynie pobieżnie. Uznaliśmy jednak za celowe wspomnienie o tych aspektach dla przedstawienia pełnego obrazu zagadnień związanych z modelowaniem systemów bezpieczeństwa. Szczegółowe opracowanie tych metod wykracza poza zakres niniejszych badań.

Planowanie w warunkach niepewności

Planowanie w warunkach niepewności uwzględnia najgorsze możliwe scenariusze w celu zapewnienia minimalnego poziomu bezpieczeństwa niezależnie od rzeczywistych wartości parametrów niepewnych [116, 117]. Podejście to optymalizuje rozwiązania pod kątem najbardziej niekorzystnych warunków, co zwiększa odporność systemu na nieprzewidywalne zmiany w środowisku operacyjnym.

4.6. Integracja z systemem COG - teoretyczne podstawy

Semantyczne mapowanie obiektów

Ontologia bezpieczeństwa

System COG umożliwia automatyczne mapowanie wykrytych elementów do ontologii bezpieczeństwa, definiującej hierarchiczne relacje między klasami obiektów [118, 119]. Formalna ontologia $O = (C, R, I)$ składa się z:

- C: Zbiór klas (concepts)
- R: Zbiór relacji między klasami
- I: Zbiór instancji (individuals) [120]

Propagacja właściwości semantycznych

Wykryte przez COG elementy dziedziczą właściwości bezpieczeństwa zgodnie ze zdefiniowaną ontologią [27]. Mechanizm inferencji umożliwia automatyczne wyprowadzenie parametrów bezpieczeństwa dla elementów o niepełnych danych.

Transformacja multimodalna

Od pikseli do grafów

Proces transformacji od surowych detekcji pikselowych do reprezentacji grafowej wymaga kilku etapów abstrakcji [48]:

1. **Segmentacja przestrzenna**: Podział obrazu na regiony semantyczne
2. **Abstrakcja topologiczna**: Reprezentacja relacji sąsiedztwa między regionami
3. **Przypisanie właściwości**: Mapowanie charakterystyk bezpieczeństwa do węzłów i krawędzi

Zachowanie spójności semantycznej

Transformacja musi zachować semantyczne właściwości oryginalnego planu architektonicznego [121]. Walidacja spójności obejmuje:

- Sprawdzenie logiki przestrzennej (np. braku "latających"pomieszczeń)
- Weryfikację ciągłości ścieżek komunikacyjnych
- Kontrolę kompletności reprezentacji elementów bezpieczeństwa

4.7. Eksploracja wielościeżkowa - podstawy teoretyczne

Teoria wielokryteriowej optymalizacji

Optymalizacja wielokryteriowa w analizie bezpieczeństwa

W analizie bezpieczeństwa często występują konfliktujące cele, które nie mogą być jednocześnie optymalizowane bez kompromisów. Typowe dylematy obejmują minimalizację czasu ataku versus minimalizację ryzyka wykrycia, maksymalizację pokrycia ochronnego versus minimalizację kosztów instalacji, czy optymalizację przepustowości ruchu versus zwiększenie poziomu kontroli bezpieczeństwa [122, 123].

Definicja optymalności Pareto: Rozwiązanie jest **Pareto-optymalne**, jeśli nie istnieje inne rozwiązanie lepsze we wszystkich kryteriach jednocześnie. Formalnie, rozwiązanie $\mathbf{x}^*$ jest Pareto-optymalne, jeśli nie istnieje $\mathbf{x}$ takie, że:

$$\forall i : f_i(\mathbf{x}) \leq f_i(\mathbf{x}^*) \wedge \exists j : f_j(\mathbf{x}) < f_j(\mathbf{x}^*)$$

gdzie $f_i(\mathbf{x})$ reprezentuje wartość i-tego kryterium dla rozwiązania $\mathbf{x}$.

Front Pareto w kontekście bezpieczeństwa: Zbiór wszystkich rozwiązań Pareto-optymalnych tworzy tak zwany front Pareto, który reprezentuje możliwe kompromisy między konfliktującymi celami. W analizie systemów ochrony front ten może ilustrować relacje między:

- **Kosztami implementacji** a **poziomem bezpieczeństwa**: wyższe nakłady finansowe zwykle przekładają się na lepszą ochronę, ale z malejącą użytecznością krańcową
- **Czasem reakcji** a **dokładnością detekcji**: szybsze systemy alarmowe mogą generować więcej fałszywych alarmów
- **Dostępnością systemu** a **restrykcyjnością kontroli**: bardziej szczelne procedury bezpieczeństwa mogą obniżać komfort użytkowania

Praktyczne zastosowania: W rzeczywistych scenariuszach wybór konkretnego rozwiązania z frontu Pareto zależy od preferencji decydenta i może być wspomagany przez metody takie jak:

- ważenie kryteriów według ich względnej ważności
- analiza wrażliwości dla różnych scenariuszy zagrożeń
- iteracyjne dostrajanie kompromisów na podstawie feedbacku operacyjnego

Takie podejście umożliwia świadome podejmowanie decyzji projektowych z pełną świadomością trade-offów między różnymi aspektami bezpieczeństwa.

Skalaryzacja funkcji celu

Metody skalaryzacji stanowią praktyczne podejście do rozwiązywania problemów wielokryteriowych poprzez przekształcenie ich w problemy jednokryteriowe. Podstawowa metoda kombinacji liniowej łączy wszystkie kryteria w jedną funkcję celu [124]:

$$F(\mathbf{x}) = \sum_{i=1}^{k} w_i \cdot f_i(\mathbf{x})$$

gdzie $w_i \geq 0$ są wagami reprezentującymi względną ważność poszczególnych kryteriów, przy czym $\sum_{i=1}^{k} w_i = 1$.

Interpretacja wag w kontekście bezpieczeństwa: Wybór wag w_i determinuje preferencje między kryteriami i umożliwia eksplorację różnych punktów na froncie Pareto. Na przykład:

- $w_1 = 0.7, w_2 = 0.3$ dla priorytetyzacji bezpieczeństwa nad kosztami
- $w_1 = 0.3, w_2 = 0.7$ dla scenariuszy o ograniczonym budżecie

Ograniczenia metody liniowej: Skalaryzacja liniowa nie może znaleźć wszystkich rozwiązań Pareto-optymalnych, szczególnie w przypadku frontu Pareto o kształcie wklęsłym. W takich sytuacjach stosuje się zaawansowane metody skalaryzacji:

Metoda ε-constraint: Jeden cel jest optymalizowany, podczas gdy pozostałe traktowane są jako ograniczenia:

$$\begin{aligned} \min_{x} \quad & f_1(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq \varepsilon_i, \quad i = 2, ..., k \end{aligned}$$

Metoda Chebysheva: Minimalizuje największe względne odchylenie od idealnego punktu:

$$\min_{\mathbf{x}} \max_{i=1,...,k} w_i |f_i(\mathbf{x}) - z_i^*|$$

Zastosowanie w systemach bezpieczeństwa: W praktyce skalaryzacja umożliwia systematyczne przeszukiwanie przestrzeni rozwiązań poprzez iteracyjną zmianę wag, co pozwala na identyfikację najbardziej odpowiednich konfiguracji systemu ochrony w zależności od specyficznych wymagań i ograniczeń obiektu.

Algorytmy k-najkrótszych ścieżek

Algorytm Yena

Algorytm Yena systematycznie znajduje k najkrótszych ścieżek między parą węzłów w grafie ważonym [125]. W kontekście analizy bezpieczeństwa umożliwia identyfikację alternatywnych tras ataku uporządkowanych według rosnących kosztów (lub malejącego prawdopodobieństwa sukcesu).

Zasada działania: Algorytm wykorzystuje podejście iteracyjne, gdzie w każdej iteracji znajduje kolejną najkrótszą ścieżkę poprzez:

1. Modyfikację grafu przez usunięcie lub zwiększenie wag krawędzi należących do wcześniej znalezionych ścieżek
2. Zastosowanie algorytmu najkrótszej ścieżki (np. Dijkstry) do zmodyfikowanego grafu
3. Przywrócenie oryginalnej struktury grafu dla następnej iteracji

Znaczenie w analizie bezpieczeństwa: Znajomość wielu alternatywnych ścieżek ataku pozwala na:

- **Priorytetyzację zabezpieczeń**: koncentrację zasobów ochronnych na węzłach krytycznych występujących w wielu trasach
- **Analizę wrażliwości**: ocenę wpływu wzmocnienia poszczególnych barier na ogólne bezpieczeństwo systemu
- **Planowanie redundancji**: identyfikację miejsc, gdzie dodatkowe zabezpieczenia przyniosą największy efekt

Interpretacja wyników: Jeśli różnica kosztów między pierwszą a drugą najkrótszą ścieżką jest mała, oznacza to, że atakujący ma wiele równoważnych opcji, co zwiększa nieprzewidywalność ataku. Duże różnice kosztów między kolejnymi ścieżkami wskazują na istnienie „wąskich gardeł"w systemie ochrony — miejsc, których wzmocnienie może znacząco poprawić ogólne bezpieczeństwo.

Ograniczenia praktyczne: Algorytm Yena ma złożoność obliczeniową $O(kN(M + N \log N))$, gdzie N to liczba węzłów, M to liczba krawędzi, a k to liczba poszukiwanych ścieżek. Dla dużych grafów reprezentujących złożone obiekty może to wymagać znacznych zasobów obliczeniowych, szczególnie przy poszukiwaniu wielu alternatywnych tras.

Algorytm Eppsteina

Algorytm Eppsteina efektywnie generuje k najkrótszych ścieżek w porządku rosnącym kosztów z złożonością $O(m + n \log n + k)$ [126]. Stanowi znaczną poprawę w stosunku do algorytmu Yena, szczególnie dla dużych grafów z potrzebą eksploracji wielu alternatywnych tras.

Kluczowe innowacje algorytmu: Algorytm Eppsteina wykorzystuje dwie fundamentalne struktury danych:

- **Graf odchyleń**: reprezentuje wszystkie możliwe odchylenia od najkrótszej ścieżki do każdego węzła
- **Kopiec ścieżek**: umożliwia efektywne generowanie kolejnych najkrótszych ścieżek w porządku rosnącym

Przewagi nad algorytmem Yena: W przeciwieństwie do algorytmu Yena o złożoności $O(kN(M + N \log N))$, algorytm Eppsteina ma złożoność praktycznie liniową względem liczby poszukiwanych ścieżek k. To sprawia, że jest szczególnie efektywny w scenariuszach wymagających analizy dużej liczby alternatywnych tras ataku.

Zastosowanie w analizie bezpieczeństwa: Efektywność algorytmu Eppsteina umożliwia:

- **Kompleksową analizę scenariuszy**: eksplorację setek potencjalnych tras ataku w rozsądnym czasie
- **Analizę w czasie rzeczywistym**: możliwość dynamicznego przeliczania alternatywnych ścieżek przy zmianie konfiguracji systemu

– **Optymalizację zasobów**: identyfikację optymalnych punktów wzmocnienia systemu ochrony na podstawie pełnej analizy przestrzeni możliwych ataków

Ograniczenia implementacyjne: Pomimo teoretycznej efektywności, praktyczna implementacja algorytmu Eppsteina wymaga znacznej złożoności programistycznej oraz dokładnego zarządzania pamięcią ze względu na konieczność utrzymywania struktur danych reprezentujących graf odchyleń dla całego systemu.

Strategia eksploracji progresywnej

Iteracyjne usuwanie krawędzi

Strategia progresywnego usuwania pierwszych krawędzi znalezionych ścieżek wymusza eksplorację alternatywnych tras poprzez systematyczne eliminowanie już przeanalizowanych opcji [127]. Teoretyczną podstawą jest twierdzenie o maksymalnym przepływie i minimalnym cięciu w grafach [128].

Mechanizm działania: Algorytm iteracyjnie modyfikuje graf poprzez:

1. Znalezienie najkrótszej ścieżki w aktualnym grafie
2. Usunięcie lub zwiększenie wagi pierwszej krawędzi tej ścieżki
3. Powtórzenie procesu dla zmodyfikowanego grafu

Uzasadnienie teoretyczne: Zgodnie z twierdzeniem Forda-Fulkersona, maksymalny przepływ w sieci równa się minimalnej wartości cięcia. W kontekście analizy bezpieczeństwa oznacza to, że usunięcie krawędzi o minimalnej wadze zmusza przeciwnika do poszukiwania alternatywnych tras o wyższych kosztach.

Zastosowanie w analizie bezpieczeństwa: Metoda umożliwia systematyczną identyfikację słabych punktów systemu poprzez:

– Odkrywanie ścieżek obchodzących główne zabezpieczenia

– Identyfikację redundantnych tras o podobnych kosztach

– Ocenę wpływu wzmocnienia poszczególnych barier na ogólną odporność systemu

Kontrola różnorodności ścieżek

Metryki różnorodności ścieżek zapewniają, że eksplorowane alternatywy są rzeczywiście odmienne, a nie jedynie wariantami podobnych tras [129]. Pozwala to uniknąć analizy redundantnych scenariuszy ataku o marginalnych różnicach.

Popularne metryki różnorodności:

– **Odległość Hamminga**: $d_H(P_1, P_2) = |\{e \in E : e \in P_1 \oplus P_2\}|$, gdzie $\oplus$ oznacza różnicę symetryczną

– **Współczynnik nakładania węzłów**: overlap$(P_1, P_2) = \frac{|V(P_1) \cap V(P_2)|}{|V(P_1) \cup V(P_2)|}$

– **Współczynnik Jaccarda dla krawędzi**: $J(P_1, P_2) = \frac{|E(P_1) \cap E(P_2)|}{|E(P_1) \cup E(P_2)|}$

Praktyczne zastosowanie: Kontrola różnorodności pozwala na:

– Unikanie analizy ścieżek o wysokim stopniu podobieństwa (np. overlap > 0.8)

– Zapewnienie reprezentatywności próbki analizowanych scenariuszy ataku

– Optymalizację czasu obliczeń poprzez eliminację redundantnych analiz

Próg różnorodności: W praktycznych zastosowaniach często ustala się minimalny próg różnorodności (np. $d_H \geq 3$ krawędzie lub overlap ≤ 0.7), poniżej którego nowe ścieżki są odrzucane jako zbyt podobne do już znalezionych tras.

4.8. Walidacja i filtracja ścieżek

Teoria ograniczeń w problemach optymalizacji

Programowanie z ograniczeniami

Walidacja ścieżek może być sformalizowana jako problem spełniania ograniczeń (CSP) [130, 131], gdzie każda potencjalna trasa ataku musi spełniać zdefiniowany zbiór warunków logicznych i fizycznych.

Formalna reprezentacja CSP dla walidacji ścieżek obejmuje:

– **Zmienne**: Węzły w ścieżce $\{v_1, v_2, ..., v_n\}$ reprezentujące kolejne lokalizacje

– **Dziedziny**: Zbiory dozwolonych wartości dla każdej zmiennej, uwzględniające ograniczenia fizyczne i logiczne

– **Ograniczenia**: Reguły określające dozwolone kombinacje wartości, takie jak:

  – Ograniczenia łączności: $connected(v_i, v_{i+1}) = true$

  – Ograniczenia czasowe: $time(v_i \rightarrow v_{i+1}) \leq threshold$

  – Ograniczenia dostępności: $accessible(v_i, context) = true$

Zastosowanie w analizie bezpieczeństwa: CSP umożliwia systematyczną eliminację nieważnych ścieżek poprzez:

– Weryfikację zgodności z ograniczeniami architektonicznymi (brak przecięć ze ścianami)

– Sprawdzenie realizowalności czasowej (czy trasa może być pokonana w dostępnym czasie)

– Walidację logiczną sekwencji działań (dostępność kluczy, uprawnień)

Propagacja ograniczeń

Algorytmy propagacji ograniczeń efektywnie eliminują niedozwolone kombinacje poprzez zapewnienie lokalnej konsystencji [132]. Proces ten znacząco redukuje przestrzeń poszukiwań poprzez wczesne wykrywanie niespełnialnych warunków.

Kluczowe techniki konsystencji:

- **Arc-consistency**: Zapewnia, że dla każdej wartości w dziedzinie zmiennej istnieje zgodna wartość w dziedzinie każdej sąsiadującej zmiennej
- **Path-consistency**: Gwarantuje, że każda para wartości zgodnych zmiennych może być rozszerzona o trzecią zmienną zachowującą spójność
- **K-consistency**: Uogólnienie zapewniające spójność dla dowolnych k-elementowych podzbiorów zmiennych

Efektywność obliczeniowa: Propagacja ograniczeń redukuje złożoność walidacji z $O(d^n)$ (gdzie d to rozmiar dziedziny, n to liczba zmiennych) do znacznie mniejszych przestrzeni poprzez przycinanie niedozwolonych kombinacji na wczesnym etapie.

Geometryczna walidacja ścieżek

Algorytmy przecięcia segmentów

Walidacja kolizji ścieżek z barierami architektonicznymi wymaga efektywnego wykrywania przecięć między segmentami reprezentującymi trasy a elementami konstrukcyjnymi [133]. Algorytm Bentley-Ottmanna stanowi klasyczne rozwiązanie tego problemu.

Algorytm Bentley-Ottmanna: Efektywnie wykrywa wszystkie przecięcia w zbiorze n segmentów z złożonością $O((n + k) \log n)$, gdzie k to liczba przecięć [134]. Algorytm wykorzystuje:

- **Linię zamiatającą**: Poruszającą się w kierunku rosnących współrzędnych x
- **Strukturę danych**: Zrównoważone drzewo przechowujące aktywne segmenty
- **Kolejkę zdarzeń**: Uporządkowane punkty początku, końca i przecięć segmentów

Adaptacja do analizy bezpieczeństwa:

- Reprezentacja barier architektonicznych jako segmentów linii
- Modelowanie ścieżek ataku jako łamanych
- Identyfikacja punktów wymagających specjalnych uprawnień lub narzędzi

Triangulacja Delaunaya

Triangulacja Delaunaya przestrzeni architektonicznej umożliwia efektywną walidację dostępności obszarów i legalności ścieżek poprzez optymalne pokrycie płaszczyzny [135]. Właściwość pustego koła gwarantuje, że żaden punkt nie leży wewnątrz okręgu opisanego na którymkolwiek trójkącie.

Zastosowania w walidacji ścieżek:

– **Planowanie ruchu**: Wierzchołki trójkątów definiują optymalne punkty nawigacyjne

– **Analiza widoczności**: Krawędzie triangulacji wyznaczają linie bez przeszkód

– **Obliczanie odległości**: Ścieżki wzdłuż krawędzi triangulacji minimalizują długość tras

Algorytm konstrukcji: Algorytm przyrostowy konstruuje triangulację w czasie $O(n \log n)$ poprzez:

1. Dodawanie punktów po kolei
2. Lokalizację trójkąta zawierającego nowy punkt
3. Aktualizację triangulacji poprzez przerzucanie krawędzi (edge flipping)

Dual graph navigation: Graf dualny triangulacji Delaunaya (diagram Voronoi) umożliwia efektywną nawigację między obszarami o różnych właściwościach bezpieczeństwa, gdzie każda komórka Voronoi reprezentuje region o jednorodnych charakterystykach ochronnych.

4.9. Integracja z metodologią EASI

Teoria niezawodności systemów

Modele szeregowo-równoległe

Systemy bezpieczeństwa można modelować jako kombinacje elementów szeregowych (gdzie awaria jednego elementu powoduje awarię całego systemu) i równoległych (z redundancją) [136, 137].

Niezawodność systemu szeregowego:

$$R_{series} = \prod_i R_i$$

Niezawodność systemu równoległego:

$$R_{parallel} = 1 - \prod_i (1 - R_i)$$

Implikacje praktyczne dla analizy bezpieczeństwa:

– **Identyfikacja słabych ogniw**: W systemach szeregowych jeden niskiej jakości element (np. $R_i = 0.7$) drastycznie obniża całkowitą niezawodność, co wskazuje na priorytetowe obszary modernizacji

– **Ocena efektywności redundancji**: Systemy równoległe wykazują szybko malejące korzyści z dodawania kolejnych elementów rezerwowych

– **Optymalizacja architektury**: Model umożliwia porównanie kosztów implementacji redundancji versus poprawy niezawodności pojedynczych komponentów

– **Planowanie konserwacji**: Krytyczne elementy w łańcuchach szeregowych wymagają częstszej kontroli i wcześniejszej wymiany

Procesy Markova w analizie degradacji

Degradacja systemu bezpieczeństwa może być modelowana jako proces Markova ze skończoną liczbą stanów [138]. Macierz przejścia $P = [p_{ij}]$ opisuje prawdopodobieństwa przejścia między stanami sprawności systemu.

Praktyczne zastosowania modeli Markova:

- **Prognozowanie awarii**: Model umożliwia przewidywanie prawdopodobieństwa osiągnięcia stanu krytycznego w określonym horyzoncie czasowym
- **Optymalizacja harmonogramów konserwacji**: Analiza czasów oczekiwanych między stanami pozwala na planowanie optymalnych interwałów serwisowych
- **Analiza kosztów eksploatacji**: Każdemu stanowi można przypisać koszty operacyjne, umożliwiając długoterminową optymalizację ekonomiczną
- **Ocena strategii utrzymania**: Porównanie różnych strategii konserwacji (proaktywna vs reaktywna) poprzez symulację macierzy przejść

Probabilistyczna analiza ścieżek ataku

Splotowe modele czasu

Czas całkowity pokonania ścieżki ataku stanowi splot czasów pokonania poszczególnych elementów [139]. Dla niezależnych składników:

$$F_{total}(t) = F_1 * F_2 * \ldots * F_n(t)$$

gdzie $*$ oznacza operację splotu, a F_i to funkcja rozkładu czasów składnikowych.

Znaczenie praktyczne analizy splotowej:

- **Analiza scenariuszy czasowych**: Model pozwala na obliczenie prawdopodobieństwa, że atak zostanie zakończony przed przybyciem sił interwencyjnych
- **Identyfikacja wąskich gardeł czasowych**: Elementy o dużej wariancji czasu pokonania wprowadzają największą niepewność do całkowitego czasu ataku
- **Optymalizacja czasu reakcji**: Określenie minimalnego czasu reakcji systemu ochrony gwarantującego skuteczną interwencję z zadanym prawdopodobieństwem
- **Planowanie ćwiczeń**: Model dostarcza realistyczne scenariusze czasowe dla szkoleń służb ochrony

Bayesowska aktualizacja parametrów

W miarę pozyskiwania nowych danych (np. z incydentów bezpieczeństwa), parametry modelu mogą być aktualizowane zgodnie z twierdzeniem Bayesa [140, 141]:

$$P(\theta|\text{data}) \propto P(\text{data}|\theta) \cdot P(\theta)$$

gdzie $P(\theta)$ to rozkład a priori parametrów, a $P(\text{data}|\theta)$ to funkcja wiarygodności.

Korzyści z podejścia bayesowskiego:

- **Adaptacyjność modelu**: System automatycznie dostosowuje swoje przewidywania na podstawie nowych obserwacji z rzeczywistych incydentów
- **Uwzględnienie niepewności**: Rozkłady a posteriori kwantyfikują niepewność oszacowań, umożliwiając podejmowanie decyzji uwzględniających ryzyko
- **Integracja wiedzy eksperckiej**: Rozkłady a priori pozwalają na włączenie doświadczenia specjalistów ds. bezpieczeństwa do modelu formalnego
- **Ciągłe doskonalenie**: Model staje się coraz dokładniejszy w miarę gromadzenia danych operacyjnych
- **Analiza wrażliwości**: Badanie wpływu różnych rozkładów a priori na wyniki końcowe pozwala na ocenę stabilności modelu

Ograniczenia i uzasadnienie wyboru metodologii

Pomimo teoretycznych zalet wymienionych podejść, w niniejszej pracy zdecydowano się na skupienie na probabilistycznym modelu degradacji CPPS z następujących powodów:

Ograniczenia modeli szeregowo-równoległych:

- **Uproszczenia strukturalne**: Rzeczywiste systemy bezpieczeństwa rzadko mają czysto szeregową lub równoległą strukturę, a ich modelowanie wymaga złożonych kombinacji hybrydowych
- **Brak uwzględnienia degradacji czasowej**: Modele zakładają stałe prawdopodobieństwa niezawodności, nie uwzględniając progresywnej degradacji komponentów
- **Nieadekwatność dla cyberzagrożeń**: Klasyczne modele niezawodności nie odzwierciedlają specyfiki ataków cybernetycznych i hybrydowych

Praktyczne ograniczenia procesów Markova:

- **Złożoność obliczeniowa**: Dla rzeczywistych systemów z dziesiątkami stanów macierze przejść stają się trudne do estymacji i interpretacji
- **Założenie braku pamięci**: Procesy Markova zakładają, że przyszłe stany zależą tylko od stanu obecnego, co nie odzwierciedla skumulowanych efektów degradacji
- **Trudności w parametryzacji**: Estymacja macierzy przejść wymaga długich serii danych historycznych, często niedostępnych dla systemów bezpieczeństwa

Wyzwania splotowych modeli czasu:

- **Założenie niezależności**: Czasy pokonania kolejnych barier często są skorelowane (np. przez poziom umiejętności atakującego)

– **Złożoność obliczeniowa**: Analityczne obliczenie splotów dla wielu rozkładów staje się niewykonalne analitycznie

– **Brak danych empirycznych**: Ograniczona dostępność danych o rzeczywistych czasach pokonania barier bezpieczeństwa

Ograniczenia podejścia bayesowskiego:

– **Subiektywność rozkładów a priori**: Wybór rozkładów a priori może znacząco wpływać na wyniki, szczególnie przy ograniczonych danych

– **Złożoność obliczeniowa**: Metody MCMC wymagają znacznych zasobów obliczeniowych dla złożonych modeli

– **Trudności w walidacji**: Ocena jakości rozkładów a posteriori w braku zewnętrznych danych referencyjnych stanowi istotne wyzwanie

W przeciwieństwie do wymienionych ograniczeń, opracowany probabilistyczny model degradacji CPPS oferuje praktyczne podejście, które:

– Uwzględnia specyfikę współczesnych zagrożeń cyber-fizycznych

– Wykorzystuje parametry możliwe do estymacji na podstawie dostępnych danych

– Zapewnia interpretowalność wyników dla praktykujących specjalistów ds. bezpieczeństwa

– Umożliwia praktyczną implementację w rzeczywistych środowiskach operacyjnych

4.10. Zastosowania w cyber-fizycznych systemach bezpieczeństwa

Teoria systemów cyber-fizycznych

Koewolucja warstw fizycznej i cybernetycznej

Cyber-fizyczne systemy bezpieczeństwa charakteryzują się ścisłą integracją komponentów fizycznych i cybernetycznych [142, 143]. Teoria koewolucji opisuje wzajemne wpływy i pętle sprzężenia zwrotnego między warstwami.

Mechanizmy koewolucji w systemach bezpieczeństwa:

– **Sprzężenie zwrotne pozytywne**: Poprawa zabezpieczeń cybernetycznych prowadzi do zwiększenia zaufania do systemów fizycznych, co z kolei motywuje dalsze inwestycje w cyberbezpieczeństwo

– **Propagacja podatności**: Słabości w warstwie cybernetycznej (np. niezaktualizowane oprogramowanie) bezpośrednio wpływają na skuteczność barier fizycznych (np. systemy kontroli dostępu)

– **Adaptacja przeciwnika**: Ewolucja metod ataku w odpowiedzi na wzmocnienia jednej warstwy prowadzi do presji na drugą warstwę systemu

Praktyczne implikacje koewolucji: Zrozumienie mechanizmów koewolucji umożliwia:

- Przewidywanie kaskadowych efektów modernizacji jednej warstwy na pozostałe komponenty
- Optymalizację inwestycji poprzez identyfikację synergii między ulepszeniami fizycznymi i cybernetycznymi
- Planowanie strategii obrony uwzględniającej dynamiczną naturę zagrożeń hybrydowych

Modele hybrydowe

Systemy hybrydowe łączą dynamikę ciągłą (fizyczną) z dyskretną (cybernetyczną) [144]. Automaty hybrydowe stanowią formalny model matematyczny opisujący takie systemy poprzez kombinację:

- **Stanów dyskretnych**: Reprezentujących konfiguracje systemu (np. tryb normalny, alarm, blokada)
- **Zmiennych ciągłych**: Opisujących parametry fizyczne (temperatura, ciśnienie, położenie)
- **Warunków przejść**: Definiujących reguły zmiany stanów na podstawie wartości zmiennych

Zastosowanie w modelowaniu bezpieczeństwa:

- **Modelowanie systemów HVAC**: Połączenie sterowania cyfrowego z dynamiką termiczną budynku
- **Systemy kontroli dostępu**: Integracja czytników biometrycznych z mechanicznymi barierami
- **Systemy alarmowe**: Fuzja danych z sensorów analogowych z logiką cyfrową podejmowania decyzji

Wyzwania modelowania hybrydowego:

- Weryfikacja właściwości bezpieczeństwa w nieskończonych przestrzeniach stanów
- Synchronizacja między zdarzeniami dyskretnymi a ewolucją ciągłą
- Obsługa niepewności w pomiarach i zakłóceń zewnętrznych

Tworzenie się właściwości systemowych

Teoria złożoności

Właściwości bezpieczeństwa często wyłaniają się z interakcji między prostymi składnikami [145]. Teoria sieci złożonych dostarcza narzędzi do analizy takich systemów poprzez metryki charakteryzujące strukturę i dynamikę sieci.

Kluczowe właściwości sieci złożonych w kontekście bezpieczeństwa:

- **Właściwości małego świata (small-world)**: Krótkie ścieżki między węzłami przy zachowaniu lokalnego klastrowania, co w systemach bezpieczeństwa oznacza szybką propagację alarmów przy lokalnej redundancji

– **Rozkłady bezskalowe (scale-free)**: Obecność węzłów o bardzo wysokim stopniu (huby), które w systemach bezpieczeństwa często pełnią rolę centralnych punktów kontroli [146]

– **Odporność i podatności systemu**: Różnicowanie odpowiedzi na losowe awarie versus celowe ataki na kluczowe węzły [147]

Metryki analizy sieci bezpieczeństwa:

– **Betweenness centrality**: Identyfikacja węzłów krytycznych dla przepływu informacji alarmowych

– **Clustering coefficient**: Ocena lokalnej redundancji i odporności na izolowane awarie

– **Network efficiency**: Pomiar ogólnej efektywności komunikacji w systemie

– **Modularity**: Wykrywanie naturalnych podsystemów i ich wzajemnych zależności

Właściwości bezpieczeństwa:

– **Kolektywna inteligencja**: Współpraca prostych sensorów prowadzi do zaawansowanych możliwości detekcji

– **Samoadaptacja**: System automatycznie dostosowuje swoją konfigurację w odpowiedzi na zmieniające się zagrożenia

– **Odporność na awarie**: Redundancja na poziomie systemu kompensuje awarie pojedynczych komponentów

Kaskadowe awarie (Cascading failures)

Modelowanie kaskadowych awarii w systemach bezpieczeństwa wymaga uwzględnienia złożonych zależności między komponentami [148]. Awarie początkowe mogą propagować się przez system, prowadząc do nieproporcjonalnych skutków.

Mechanizmy propagacji awarii:

– **Przeciążenie funkcjonalne**: Awaria jednego komponentu prowadzi do przekierowania obciążenia na elementy sąsiednie

– **Zależności logiczne**: Awaria systemu nadrzędnego powoduje niefunkcjonalność komponentów zależnych

– **Efekty kaskadowe w czasie**: Opóźnione reakcje systemu na pierwotną awarię mogą wyzwalać kolejne fale awarii

Grafy zależności i sieci multipleks: Współczesne narzędzia analizy obejmują:

– **Dependency graphs**: Modelowanie bezpośrednich zależności funkcjonalnych między komponentami

– **Multiplex networks**: Reprezentacja wielu typów relacji (komunikacyjne, energetyczne, logiczne) w jednym frameworku

– **Interdependent networks**: Analiza wzajemnych zależności między odrębnymi sieciami (np. IT a systemy fizyczne)

Strategie mitygacji kaskadowych awarii:

– **Dywersyfikacja zależności**: Unikanie koncentracji krytycznych funkcji w pojedynczych węzłach

– **Circuit breakers**: Automatyczne mechanizmy izolacji zapobiegające propagacji awarii

– **Graceful degradation**: Projektowanie systemów umożliwiających częściową funkcjonalność przy awarii komponentów

– **Adaptive reconfiguration**: Dynamiczne przeprogramowanie sieci w odpowiedzi na wykryte awarie

Modelowanie probabilistyczne kaskad:

– Macierze prawdopodobieństw propagacji między parami węzłów

– Modele progowe określające warunki przełączenia stanów komponentów

– Symulacje Monte Carlo dla oceny rozkładów skutków potencjalnych awarii inicjujących

4.11. Podsumowanie teoretyczne

Przedstawiona analiza teoretyczna ustanawia dobre podstawy dla algorytmicznego podejścia do analizy bezpieczeństwa opartego na teorii grafów i heurystycznym przeszukiwaniu. Przegląd literatury ujawnił znaczące luki w istniejących metodologiach, szczególnie w kontekście współczesnych zagrożeń cyber-fizycznych, co skierowało badania w kierunku hybrydowego podejścia łączącego klasyczne metody z nowoczesnymi technikami.

Wybór kierunku rozwoju metodologii oceny efektywności systemów

Na podstawie przeprowadzonej analizy teoretycznej zdecydowano się na **wzmocnienie metodologii EASI** poprzez integrację trzech kluczowych elementów:

1. **Rozszerzenie o zagrożenia z cyberprzestrzeni**: Klasyczna metodologia EASI, opracowana w erze czysto fizycznych systemów ochrony, wymaga adaptacji do rzeczywistości cyber-fizycznej. Zintegrowano elementy modelowania ataków zdalnych, podatności systemów IT oraz hybrydowych scenariuszy zagrożeń.

2. **Wprowadzenie dynamicznego modelu degradacji**: Tradycyjne założenie o statycznych parametrach skuteczności barier zostało zastąpione probabilistycznym modelem degradacji uwzględniającym starzenie się komponentów, wzrost wiedzy przeciwnika, postęp technologiczny oraz działania konserwacyjne.

3. **Automatyzację procesu analizy**: Manualny charakter tradycyjnej analizy EASI ogranicza jej skalowalność i spójność. Zaimplementowano detekcję elementów architektonicznych opartą o ML (YOLO) oraz algorytmiczną eksplorację ścieżek ataku (własna funkcja heurystyki dla A*).

Uzasadnienie wyborów metodologicznych

Decyzja o rozszerzeniu EASI, zamiast adaptacji alternatywnych podejść (teoria gier, optymalizacja bayesowska, modele Markova), wynikała z następujących przesłanek:

- **Zachowanie kompatybilności**: EASI pozostaje standardem w branży bezpieczeństwa, co zapewnia akceptację i łatwość adaptacji rozwiązania
- **Praktyczna implementowalność**: Proponowane rozszerzenia można wdrożyć stopniowo, nie wymagając rewolucyjnych zmian w istniejących procesach
- **Dostępność danych**: Model degradacji wykorzystuje parametry możliwe do estymacji na podstawie dostępnych danych eksploatacyjnych
- **Interpretowalność wyników**: Zachowanie podstawowej struktury EASI zapewnia zrozumiałość wyników dla praktykujących specjalistów

Wkład w teorię analizy bezpieczeństwa

Opracowane podejście wnosi następujące elementy do teorii analizy bezpieczeństwa:

1. **Formalizację reprezentacji grafowej** systemów bezpieczeństwa z uzasadnieniem wyboru grafów skierowanych dla precyzyjnego modelowania asymetrii bezpieczeństwa i jednokierunkowych przepływów w systemach kontroli dostępu i ścieżek ataku.
2. **Rozwój specjalizowanych funkcji heurystycznych** dostosowanych do specyfiki scenariuszy bezpieczeństwa, wykraczających poza klasyczne metryki geometryczne i uwzględniających czynniki takie jak prawdopodobieństwo wykrycia, czas reakcji oraz dostępność zasobów, czy przygotowanie adwersarza.
3. **Integrację teorii grafów z systemami percepcji wizualnej** poprzez semantyczne mapowanie wykrytych elementów architektonicznych na właściwości bezpieczeństwa, umożliwiające automatyczne generowanie modeli grafowych z planów 2D.
4. **Teoretyczne podstawy dla wielokryteriowej eksploracji ścieżek** uwzględniającej różne profile przeciwników i scenariusze zagrożeń poprzez adaptacyjne funkcje kosztu i mechanizmy kontroli różnorodności potencjalnych tras reprezentujących ścieżki ataku.
5. **Rozszerzenie klasycznych metodologii oceny bezpieczeństwa** o aspekty degradacji czasowej i cyberodporności poprzez probabilistyczny model degradacji CPPS integrujący pięć kluczowych czynników wpływających na długoterminową skuteczność systemów.

6. **Metodologia dla hybrydowej analizy zagrożeń** łączący tradycyjne modele fizycznego przełamywania barier z współczesnymi scenariuszami cyber-ataków, uwzględniający specyfikę ataków zdalnych i ich wpływ na systemy kontroli fizycznej.

Znaczenie dla praktyki

Przedstawione podejście teoretyczne może stanowić fundament dla przejścia od tradycyjnych, eksperckich metod analizy bezpieczeństwa do zautomatyzowanych, skalowalnych i konsystentnych systemów oceny zagrożeń w cyber-fizycznych środowiskach bezpieczeństwa. Kluczowe korzyści praktyczne obejmują:

- Redukcję subiektywności analiz poprzez formalizację procesów oceny
- Zwiększenie efektywności poprzez automatyzację czasochłonnych procesów manualnych
- Poprawę jakości decyzji poprzez uwzględnienie dynamicznej natury zagrożeń
- Umożliwienie analizy w skali niemożliwej do osiągnięcia metodami tradycyjnymi
- Umożliwienie racjonalnej implementacji dodatkowych środków mitygacji zagrożeń pochodzących zarówno z obszaru fizycznego jak i cyberprzestrzeni.

Opracowana metodologia stanowi pomost między dorobkiem teoretycznym a potrzebami praktyki, oferując narzędzia dostosowane do rzeczywistości współczesnych systemów bezpieczeństwa charakteryzujących się rosnącą złożonością i integracją domeny cyber-fizycznej.

5. Zastosowanie dużych modeli językowych w konsolidacji i analizie danych bezpieczeństwa

W poprzednim rozdziale omówiono formalizację wiedzy o infrastrukturze chronionego obiektu w postaci grafów ścieżek ataku oraz heurystyczne metody ich oceny. Podejście to pozwala w sposób systematyczny analizować potencjalne scenariusze działań przeciwnika, lecz generuje dużą ilość danych i wyników cząstkowych, których interpretacja wymaga zaawansowanego wsparcia. W tym celu postanowiliśmy użyć właściwości dużych modeli językowych (LLM), które dzięki zdolności do syntezy informacji i uogólniania wniosków mogą wspierać proces raportowania, rekomendacji oraz integracji wiedzy z różnych źródeł. Rozdział ten przedstawia wykorzystanie LLM jako kolejnego poziomu wspomagania decyzji, łączącego wyniki analiz probabilistycznych i topologicznych z praktycznym wymiarem zarządzania bezpieczeństwem organizacji.

5.1. Duże modele językowe - podstawy teoretyczne

Definicja i charakterystyka LLM

Duże modele językowe (Large Language Models, LLM) stanowią klasę modeli uczenia maszynowego opartych na architekturze transformera, charakteryzujących się znaczną liczbą parametrów (od miliardów do bilionów) oraz zdolnością do przetwarzania i generowania tekstu w języku naturalnym [149, 150]. Kluczowe cechy LLM obejmują:

- **Skala parametrów**: Współczesne modele jak GPT-4 czy Claude zawierają setki miliardów parametrów, umożliwiając reprezentację złożonych wzorców językowych

- **Pretrening na dużych korpusach**: Modele trenowane są na rozległych zbiorach tekstów z internetu, książek i innych źródeł, obejmujących tryliony tokenów

- **Zdolności**: Przy przekroczeniu pewnej skali pojawiają się nieoczekiwane zdolności, takie jak rozumowanie wieloetapowe czy rozwiązywanie problemów matematycznych

- **Uniwersalność zastosowań**: Jeden model może wykonywać różnorodne zadania językowe bez specjalistycznego treningu

Architektura transformer i mechanizm uwagi

Architektura transformer [149] stanowi fundament współczesnych LLM, wykorzystując mechanizm uwagi (attention) do przetwarzania sekwencji tokenów:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

gdzie Q (queries), K (keys) i V (values) to macierze reprezentujące różne aspekty danych wejściowych, a d_k to wymiarowość klucza.

Kluczowe komponenty architektury:

- **Multi-head attention**: Równoległe przetwarzanie różnych aspektów relacji między tokenami
- **Feed-forward networks**: Nieliniowe transformacje w każdej warstwie
- **Residual connections**: Połączenia resztkowe ułatwiające trenowanie głębokich sieci
- **Layer normalization**: Stabilizacja procesu uczenia

Tokenizacja i reprezentacja tekstu

Proces przetwarzania tekstu w LLM rozpoczyna się od tokenizacji - dzielenia tekstu na podstawowe jednostki zwane tokenami [151]:

Rodzaje tokenizacji:

- **Word-level**: Każde słowo jako osobny token (rzadko używane ze względu na wielkość słownika)
- **Character-level**: Każdy znak jako token (nieefektywne dla długich sekwencji)
- **Subword tokenization**: Podział na fragmenty słów, najpopularniejsze podejścia:
  - *Byte Pair Encoding (BPE)*: Iteracyjne łączenie najczęstszych par znaków
  - *SentencePiece*: Unifikowane podejście traktujące spacje jako zwykłe znaki
  - *WordPiece*: Wykorzystywane w modelach BERT i pochodnych

Właściwości tokenów:

- Jeden token odpowiada średnio 3-4 znakom w języku angielskim
- Słowniki współczesnych modeli zawierają 50,000-100,000 tokenów
- Tokeny specjalne oznaczają początek/koniec sekwencji, nieznane słowa
- Encoding pozycyjny dodaje informację o kolejności tokenów w sekwencji

Proces treningu i fine-tuningu

Współczesne LLM przechodzą wieloetapowy proces trenowania:

Pretrening (unsupervised):

- Przewidywanie następnego tokena w sekwencji (autoregresyjna funkcja straty)
- Trenowanie na rozległych, nieoznaczonych korpusach tekstowych
- Wykorzystanie ogromnych zasobów obliczeniowych (tysiące GPU przez miesiące)
- Funkcja straty: $L = -\sum_{i=1}^{n} \log P(w_i|w_1, ..., w_{i-1})$

Fine-tuning (supervised):

- Dostrajanie modelu na specjalistycznych zbiorach danych
- Instruction tuning - uczenie wykonywania instrukcji w języku naturalnym
- Reinforcement Learning from Human Feedback (RLHF) - optymalizacja zgodnie z preferencjami ludzkimi
- Constitutional AI - trenowanie zgodne z określonymi zasadami etycznymi

Zdolności i ograniczenia LLM

Zdolności LLM:

- **Few-shot learning**: Wykonywanie zadań na podstawie kilku przykładów w kontekście
- **Chain-of-thought reasoning**: Rozwiązywanie problemów przez eksplicytne przedstawienie kroków rozumowania
- **Transfer learning**: Zastosowanie wiedzy z jednej domeny do innych
- **Multimodalne możliwości**: Integracja tekstu z obrazami, dźwiękiem (w najnowszych modelach)

Istotne ograniczenia:

- **Brak dostępu do aktualnych informacji**: Wiedza ograniczona do daty granicznej zbioru treningowego
- **Halucynacje**: Generowanie wiarygodnie brzmiących, ale nieprawdziwych informacji
- **Ograniczenia kontekstowe**: Maksymalna długość przetwarzanej sekwencji (obecnie 32k-200k tokenów)
- **Brak trwałej pamięci**: Każda konwersacja rozpoczyna się od nowa
- **Nieprzewidywalność**: Trudność w zagwarantowaniu deterministycznych wyników

Zastosowania w analizie danych i automatyzacji

Modele LLM znajdują szerokie zastosowanie w analizie i przetwarzaniu informacji:

- **Przetwarzanie dokumentów**: Ekstrakcja informacji, streszczanie, analiza sentymentu
- **Generacja kodu**: Automatyczne pisanie i debugowanie programów
- **Analiza danych**: Interpretacja wyników, generowanie raportów
- **Wspomaganie decyzji**: Analiza scenariuszy, identyfikacja wzorców
- **Interfejsy konwersacyjne**: Umożliwienie interakcji z systemami w języku naturalnym

W kontekście analizy bezpieczeństwa, LLM mogą wspierać procesy oceny zagrożeń poprzez automatyczną interpretację planów, identyfikację potencjalnych słabości oraz generowanie scenariuszy testowych, co stanowi przedmiot dalszej analizy w niniejszej pracy.

5.2. Teoretyczne podstawy integracji LLM w automatyzacji oceny systemów bezpieczeństwa

Automatyczna detekcja elementów infrastruktury, konstrukcja grafów reprezentujących racjonalne ścieżki ataku oraz heurystyczna ocena z wykorzystaniem algorytmu A* generują rozległe, wielowymiarowe zbiory danych charakteryzujące profile zagrożeń. Transformacja tych danych technicznych w praktyczne wnioski strategiczne stanowi klasyczny problem przejścia od analizy ilościowej do syntezy jakościowej – obszar, w którym duże modele językowe (LLM) wykazują szczególne kompetencje [150, 152].

Z perspektywy reprezentacji informacji, problem ten można sformalizować jako transformację:

$$T : D_{technical} \times C_{context} \rightarrow R_{strategic}$$

gdzie:

- $D_{technical}$ reprezentuje skonsolidowane dane z analizy ścieżek, parametry systemów bezpieczeństwa i charakterystyki infrastruktury
- $C_{context}$ obejmuje kontekst zewnętrzny (geopolityczny, środowiskowy, operacyjny)
- $R_{strategic}$ stanowi zestaw rekomendacji strategicznych z priorytetyzacją i analizą cost-benefit

Paradygmat przejścia od fragmentarycznej do holistycznej analizy

Tradycyjne podejścia do oceny systemów bezpieczeństwa charakteryzują się fragmentarycznością, czyli oddzielną analizą poszczególnych komponentów (sensory, bariery, procedury) z późniejszą próbą ich integracji przez ekspertów. Taka metodologia prowadzi do sub-optymalizacji lokalnej, gdzie optymalizacja pojedynczych elementów nie gwarantuje optymalności systemu jako całości [7].

Właściwości systemów bezpieczeństwa

Systemy bezpieczeństwa fizycznego wykazują właściwości wynikające z interakcji między komponentami, które nie są widoczne przy analizie izolowanych elementów. LLM, dzięki zdolności do przetwarzania kontekstu o znacznej rozpiętości, mogą identyfikować te wyłaniające się wzorce poprzez:

Analizę korelatywną: Identyfikację ukrytych zależności między pozornie niezależnymi elementami systemu, takimi jak wpływ degradacji jednego sensora na efektywność całego łańcucha detekcji.

Kontekstową interpretację: Uwzględnienie czynników zewnętrznych (wydarzenia geopolityczne, zmiany w metodach operacyjnych przeciwników) w ocenie lokalnych parametrów bezpieczeństwa.

Temporalną syntezę: Integrację statycznych parametrów systemów z dynamicznymi modelami degradacji czasowej w celu przewidywania ewolucji profilu zagrożeń.

Teoria złożoności obliczeniowej w kontekście analizy bezpieczeństwa

Problem kompleksowej analizy systemów bezpieczeństwa można sklasyfikować jako należący do klasy NP-trudnych ze względu na kombinatoryczną eksplozję możliwych scenariuszy ataku oraz nieliniowe interakcje między zmiennymi systemu. LLM oferują aproksymacyjne rozwiązanie tego problemu poprzez wykorzystanie wzorców nauczonych na rozległych korpusach tekstu zawierających wiedzę ekspercką.

Z perspektywy teorii aproksymacji, LLM można interpretować jako funkcję $f_{LLM} : \mathbb{R}^n \to \mathbb{R}^m$ aproksymującą niedostępną analitycznie funkcję eksperckiej oceny f_{expert}, gdzie jakość aproksymacji zależy od:

- Reprezentatywności danych treningowych względem domeny bezpieczeństwa
- Adekwatności architektury modelu do problemu wielokryteriowej optymalizacji
- Efektywności mechanizmów prompt engineering w przekazywaniu kontekstu domenowego

5.3. Teoretyczne podstawy strukturyzowanej inżynierii promptów w domenie bezpieczeństwa

Efektywne wykorzystanie dużych modeli językowych w analizie systemów bezpieczeństwa wymaga systematycznego podejścia do konstrukcji promptów, które przekracza proste zapytania doraźne. Inżynieria promptów w tym kontekście stanowi proces przekształcania nietekstowych danych technicznych (wyniki A*, parametry COG, dane degradacyjne) w strukturyzowane zapytania umożliwiające LLM generowanie wartościowych spostrzeżeń analitycznych [153, 154].

Teoria strukturyzowanego prompt engineering

Klasyczne podejście do inżynierii promptów opiera się na iteracyjnym dostrajaniu zapytań przez ekspertów domenowych. Jednak w kontekście analizy bezpieczeństwa, gdzie spójność i powtarzalność są krytyczne, wymagane jest bardziej systematyczne podejście oparte na formalnych szablonach. Tę tzw. strukturę "template-based prompt engineering"można sformalizować jako funkcję kompozycji:

$$P_{\text{końcowy}} = \mathcal{T}(S_{struktura}, D_{domena}, C_{kontekst})$$

gdzie:

- $S_{struktura}$ reprezentuje meta-strukturę analityczną (sekwencja pytań, poziomy abstrakcji, relacje przyczynowo-skutkowe)
- D_{domena} oznacza specjalistyczną wiedzę domenową (terminologia bezpieczeństwa, standardy branżowe, typowe wzorce zagrożeń)
- $C_{kontekst}$ obejmuje dane kontekstowe specyficzne dla analizowanego przypadku
- $\mathcal{T}$ jest operatorem kompozycji szablonów implementowanym przez bibliotekę Outlines

Optymalizacja zapytań z użyciem strukturyzowanej generacji promptów na przykładzie biblioteki „Outlines”

Biblioteka Outlines wprowadza fundamentalną zmianę w podejściu do konstrukcji promptów poprzez implementację **structured generation** - mechanizmu wymuszającego określone formaty wyjściowe bezpośrednio na poziomie procesu generowania tekstu przez LLM [155]. Podejście to eliminuje tradycyjne ograniczenia związane z nieprzewidywalnością formatów wyjściowych oraz koniecznością kosztownego post-processingu, oferując bezpośrednią kontrolę nad strukturą generowanych analiz bezpieczeństwa.

Teoretyczne podstawy structured generation

Tradycyjne podejścia do formatowania wyjścia LLM opierają się na post-processingu, czyli generacji swobodnego tekstu z późniejszym parsowaniem i walidacją. Metodologia ta wprowadza jednak znaczące ryzyko utraty informacji oraz generuje dodatkowe koszty obliczeniowe związane z wieloetapowym przetwarzaniem. Outlines implementuje alternatywne podejście w postaci generacji strukturyzowanej, gdzie ograniczenia formatu są egzekwowane w trakcie procesu generowania na poziomie ’logitów’, zapewniając jednocześnie efektywność obliczeniową oraz deterministyczność wyników.

Matematycznie, standardowy proces generowania można opisać jako:

$$P(x_t|x_{1:t-1}) = \text{softmax}(f(x_{1:t-1}))$$

gdzie f reprezentuje forward pass modelu językowego. Outlines modyfikuje ten proces, wprowadzając **constraint function** C:

$$P_{constrained}(x_t|x_{1:t-1}) = \text{softmax}(C(f(x_{1:t-1})))$$

gdzie C zeruje prawdopodobieństwa tokenów naruszających określone constraints, efektywnie kierując proces generowania w stronę pożądanych struktur wyjściowych.

Hierarchiczna organizacja wiedzy eksperckiej Biblioteka "outlines"umożliwia implementację hierarchicznej struktury wiedzy eksperckiej poprzez zagnieżdżone szablony i logikę warunkową, co stanowi istotny postęp w systematyzacji procesów analitycznych. Ta architektura odzwierciedla naturalną organizację eksperckiego myślenia o bezpieczeństwie, gdzie różne poziomy abstrakcji wymagają różnych typów rozumowania oraz dostępu do specyficznych zbiorów wiedzy domenowej.

Na poziomie strategicznym funkcjonują ogólne zasady oceny ryzyka, wymagania zgodności regulacyjnej oraz analiza wpływu incydentów bezpieczeństwa na ciągłość działalności organizacji. Ten poziom charakteryzuje się wysokim stopniem abstrakcji oraz orientacją na długoterminowe konsekwencje decyzji bezpieczeństwa. Poziom taktyczny obejmuje specyficzne dla typu obiektu modele zagrożeń, typowe wektory ataku charakterystyczne dla danej branży oraz standardowe środki przeciwdziałania wypracowane przez praktykę ekspercką. Poziom operacyjny koncentruje się na szczegółowych parametrach technicznych systemów bezpieczeństwa, specyfikacjach konfiguracji sprzętowych oraz precyzyjnych procedurach konserwacji i utrzymania.

Hierarchiczność ta pozwala na adaptacyjne generowanie promptów dostosowanych do wymaganego poziomu abstrakcji oraz automatyczne zarządzanie złożonością kontekstu w zależności od ograniczeń tokenowych, co jest kluczowe dla ekonomicznej efektywności systemu w środowiskach produkcyjnych.

Budowanie szablonu i jego modularność Biblioteka "outlines"implementuje zaawansowane mechanizmy **template composition**, umożliwiające składanie złożonych promptów z modularnych komponentów w sposób analogiczny do zasad inżynierii oprogramowania. Każdy moduł może być niezależnie testowany, walidowany i optymalizowany, co prowadzi do bardziej niezawodnych i zarządzalnych systemów analitycznych, jednocześnie umożliwiając iteracyjne doskonalenie poszczególnych aspektów analizy bez wpływu na całość systemu.

Proces kompozycji można sformalizować jako:

$$\mathcal{T}_{composite} = \mathcal{T}_1 \circ \mathcal{T}_2 \circ \ldots \circ \mathcal{T}_n$$

gdzie każdy $\mathcal{T}_i$ reprezentuje specjalizowany szablon odpowiedzialny za konkretny aspekt analizy (detekcja anomalii, ocena ryzyka, generowanie rekomendacji). Taka modularność umożliwia nie tylko efektywne zarządzanie złożonością, ale również elastyczne dostosowywanie procesu analitycznego do specyficznych wymagań różnych typów obiektów lub scenariuszy bezpieczeństwa.

5.4. Wykrywanie wzorców i anomalii jako właściwość LLM

Teoretyczne podstawy rozpoznawania wzorców w LLM

Duże modele językowe wykazują zdolności do identyfikacji wzorców, które nie były bezpośrednio obecne w danych treningowych. W kontekście analizy bezpieczeństwa ta właściwość może prowadzić do odkrycia nieoczywistych korelacji między pozornie niezależnymi parametrami systemu [152, 156].

Mechanizm ten można teoretycznie opisać jako proces niejawnej ekstrakcji cech, gdzie model identyfikuje ukryte reprezentacje w przestrzeni wysokowymiarowej:

$$f_{wzorce} : \mathbb{R}^n \to \mathbb{R}^k, \quad k \ll n$$

gdzie wysokowymiarowa przestrzeń wejściowa $\mathbb{R}^n$ (zawierająca wszystkie parametry techniczne systemu) jest mapowana do niskowymiarowej przestrzeni wzorców $\mathbb{R}^k$, w której wyłaniające się wzorce stają się identyfikowalne.

Detekcja anomalii jako produkt uboczny analizy kontekstowej

Modele LLM mogą identyfikować anomalie nie poprzez bezpośrednią analizę statystyczną, lecz jako konsekwencję kontekstowego rozumowania względem oczekiwanych wzorców w domenie bezpieczeństwa. Ten mechanizm różni się znacząco od tradycyjnych metod detekcji anomalii:

Kluczowe różnice w podejściu:

- **Semantyczne vs. statystyczne**: Tradycyjne metody opierają się na odchyleniach statystycznych od normy, podczas gdy LLM rozpoznają anomalie na podstawie semantycznego znaczenia i logicznej spójności
- **Kontekst vs. izolacja**: Klasyczne algorytmy analizują dane punktowo, LLM uwzględniają szeroki kontekst i wzajemne relacje między elementami
- **Adaptacyjność**: LLM mogą dostosowywać definicję "normalności"do specyfiki danego środowiska bez konieczności ponownego treningu
- **Wielomodalność**: Zdolność do integracji różnych typów informacji (tekst, metadane, kontekst czasowy) w jednym procesie analizy

Kategorie anomalii wykrywanych przez LLM:

Anomalia kontekstowa: Element może być statystycznie normalny, ale odstający od normy w określonym kontekście operacyjnym (np. sensor o standardowych parametrach umieszczony w nietypowej lokalizacji względem procedur bezpieczeństwa lub logiki architektury budynku)

Anomalia czasowa: Konfiguracja może być optymalna obecnie, ale anomalna w perspektywie długoterminowej degradacji systemu lub zmieniających się wzorców zagrożeń (np. system skuteczny przeciwko tradycyjnym intruzom, ale podatny na nowoczesne techniki cyber-fizyczne)

Anomalia międzydziedzinowa: Wzorzec może być normalny w dziedzinie bezpieczeństwa fizycznego, ale anomalny, gdy uwzględnia się kontekst geopolityczny, regulacyjny lub technologiczny (np. standardowe zabezpieczenia nieadekwatne do aktualnego poziomu zagrożenia)

Anomalia logiczna: Konfiguracje spełniające formalne wymagania, ale naruszające logikę bezpieczeństwa lub dobre praktyki (np. systemy zapasowe umieszczone w tej samej strefie ryzyka co systemy główne)

Anomalia emergentna: Właściwości anomalne wyłaniające się z interakcji między normalnie funkcjonującymi komponentami (np. interferencja między systemami, które indywidualnie działają prawidłowo)

Przewagi nad metodami klasycznymi:

- **Redukcja false positives**: Lepsze rozumienie kontekstu zmniejsza liczbę fałszywych alarmów
- **Wykrywanie anomalii "zero-day"**: Identyfikacja nowych typów zagrożeń niewystępujących w danych treningowych
- **Wyjaśnialność**: Możliwość generowania zrozumiałych wyjaśnień dlaczego coś zostało uznane za anomalię
- **Adaptacja do domeny**: Automatyczne dostosowanie do specyfiki konkretnego obiektu bez manualnej konfiguracji

Nowe perspektywy analityczne wykraczające poza ekspercką intuicję

Kluczową wartością teoretyczną analizy opartej na LLM jest potencjał do generowania spostrzeżeń, które wykraczają poza tradycyjne wzorce eksperckiego myślenia. Mechanizm ten można skonceptualizować jako:

$$I_{nowe} = \mathcal{L}(D_{kompleksowe}) \setminus \mathcal{E}(D_{specjalistyczne})$$

gdzie $\mathcal{L}(D_{kompleksowe})$ reprezentuje zbiór spostrzeżeń generowanych przez LLM na podstawie kompleksowego kontekstu danych, a $\mathcal{E}(D_{specjalistyczne})$ oznacza spostrzeżenia dostępne tradycyjnej analizie eksperckiej skupionej na określonych obszarach specjalistycznych.

Ta różnica może wynikać z:

- **Transferu wiedzy międzydziedzinowej:** LLM może stosować wzorce z innych dziedzin (cyberbezpieczeństwo, planowanie miejskie, zarządzanie kryzysowe) do bezpieczeństwa fizycznego
- **Skalowalnego rozpoznawania wzorców:** Zdolność do jednoczesnej analizy wszystkich parametrów systemu bez ograniczeń poznawczych wpływających na ekspertów ludzkich
- **Integracji perspektywy czasowej:** Naturalne uwzględnienie długoterminowych trendów degradacyjnych w krótkoterminowych rekomendacjach taktycznych

5.5. Konsolidacja i strukturyzacja danych dla potrzeb analitycznych

Teoretyczne podstawy konsolidacji danych

Problem konsolidacji heterogenicznych danych pochodzących z różnych etapów automatycznej analizy stanowi przypadek szczególny szerszej teorii fuzji informacji [157], wymagający zaadoptowania klasycznych podejść do specyfiki domeny bezpieczeństwa. Proces ten obejmuje integrację wyników z systemów detekcji obiektów YOLO, konstrukcji modeli grafowych, algorytmów wyszukiwania ścieżek A*, oraz mapowania kontekstowego COG, gdzie każdy komponent generuje dane o różnej naturze, szczegółowości oraz poziomie abstrakcji semantycznej.

Rozwiązywanie luki semantycznej stanowi znaczne wyzwanie teoretyczne w procesie konsolidacji, wymagające systematycznej transformacji numerycznych identyfikatorów w znaczące etykiety semantyczne dostępne dla analizy eksperta. Problem ten wykracza poza proste mapowanie jeden-do-jeden, ponieważ często wymaga kontekstualnej interpretacji, gdzie ten sam numeryczny identyfikator węzła grafu może reprezentować różne elementy architektoniczne w zależności od jego pozycji topologicznej, relacji z sąsiadującymi węzłami oraz charakterystyki funkcjonalnej w ramach szerszego systemu bezpieczeństwa. Proces ten wymaga opracowania zaawansowanych ontologii domenowych oraz mechanizmów wnioskowania kontekstualnego, które mogą skutecznie przełożyć abstrakcyjne reprezentacje matematyczne na intuicyjne reprezentacje eksperckie.

Wyzwanie wyrównania szczegółowości dotyczy istotnej różnicy między szczegółowością danych technicznych generowanych przez systemy automatyczne a poziomem abstrakcji wymaganym do strategicznego podejmowania decyzji przez kierownictwo organizacji. Systemy automatyczne mogą generować tysiące precyzyjnych parametrów technicznych dotyczących konkretnych elementów infrastruktury, podczas gdy decydenci potrzebują skonsolidowanych, wysokopoziomowych wskaźników informujących o ogólnym stanie bezpieczeństwa oraz priorytetach działań naprawczych. Teoretyczne podejście do tego problemu wymaga opracowania hierarchicznych struktur agregacji, które mogą skutecznie redukować wymiarowość danych przy jednoczesnym zachowaniu krytycznej informacji niezbędnej do podejmowania świadomych decyzji strategicznych.

Zarządzanie propagacją niepewności przez kolejne etapy potoku analitycznego stanowi jeden z najbardziej złożonych aspektów teoretycznych całego procesu. Każdy etap automatycznej analizy wprowadza własne źródła niepewności - od błędów klasyfikacji w systemach YOLO, przez nieprecyzyjności w konstrukcji grafów, po heurystyczne przybliżenia w algorytmach wyszukiwania ścieżek. Tradycyjne podejścia do propagacji niepewności, oparte na teorii prawdopodobieństwa czy teorii zbiorów rozmytych, mogą nie być adekwatne do specyfiki nieliniowych zależności charakterystycznych dla systemów bezpieczeństwa. Konieczne może być opracowanie specjalizowanych metodologii uwzględniających korelacje między różnymi źródłami niepewności oraz ich nieaddytywny charakter w kontekście oceny ryzyka bezpieczeństwa.

Model hierarchicznej konsolidacji danych

Efektywna konsolidacja danych w kontekście analizy systemów bezpieczeństwa wymaga przyjęcia hierarchicznej struktury uwzględniającej różne poziomy abstrakcji analitycznej, gdzie każdy poziom służy określonej funkcji w łańcuchu wartości analitycznej oraz adresuje specyficzne potrzeby różnych grup interesariuszy.

Warstwa danych surowych stanowi fundament całej struktury analitycznej, gromadząc nieprzetworzone wyjścia z systemów automatycznej detekcji bez jakiejkolwiek interpretacji kontekstowej. Na tym poziomie funkcjonują podstawowe dane techniczne, takie jak ramki ograniczające obiekty wykryte przez systemy wizyjne, współrzędne węzłów w modelach grafowych reprezentujących topologię budynku, czy wagi krawędzi odzwierciedlające koszty poruszania się między różnymi elementami infrastruktury. Kluczową charakterystyką tej warstwy jest jej neutralność semantyczna - dane pozostają w postaci nume-

rycznej, bez przypisanej interpretacji domenowej, co zapewnia elastyczność w późniejszym przetwarzaniu oraz umożliwia wielokrotne wykorzystanie tych samych danych bazowych w różnych kontekstach analitycznych.

Warstwa semantyczna wprowadza transformacje danych surowych w informacje znaczące poprzez mapowanie na specjalizowane konteksty domenowe. Proces ten obejmuje przypisanie semantycznych etykiet do wykrytych obiektów (identyfikacja konkretnych typów elementów architektonicznych), klasyfikację funkcji bezpieczeństwa poszczególnych komponentów systemu ochrony oraz ocenę istotności identyfikowanych zagrożeń w kontekście specyficznych scenariuszy operacyjnych. Ta warstwa stanowi kluczowy mostek między technicznymi możliwościami systemów automatycznych a ekspercką wiedzą domenową, umożliwiając skuteczną komunikację między komponentami technologicznymi a analitykami bezpieczeństwa.

Warstwa analityczna koncentruje się na agregacji i wyprowadzaniu wysokopoziomowych metryk oraz spostrzeżeń analitycznych z danych semantycznych. Na tym poziomie generowane są wskaźniki podatności systemu na różne typy ataków, identyfikowane są luki w pokryciu monitoringowym czy fizycznym, oraz kalkulowane współczynniki efektywności kosztowej różnych opcji modernizacyjnych. Proces analityczny może wykorzystywać zarówno deterministyczne algorytmy, jak i probabilistyczne modele, w zależności od charakteru analizowanych problemów oraz dostępności danych historycznych. Kluczowym aspektem tej warstwy jest jej zdolność do syntezy fragmentarycznych obserwacji w koherentny obraz stanu bezpieczeństwa.

Warstwa strategiczna reprezentuje najwyższy poziom abstrakcji, transformując wyniki analityczne w rekomendacje wysokiego poziomu oraz podsumowania wykonawcze dostosowane do potrzeb podejmowania decyzji przez kierownictwo organizacji. Na tym poziomie techniczne szczegóły są abstrahowane w zestaw klarownych wskazówek, które mogą bezpośrednio informować procesy planowania strategicznego, alokacji budżetu bezpieczeństwa oraz priorytetyzacji inicjatyw modernizacyjnych. Warstwa ta musi uwzględniać nie tylko techniczne aspekty bezpieczeństwa, ale również ograniczenia organizacyjne, regulacyjne oraz finansowe charakterystyczne dla konkretnego kontekstu operacyjnego.

Optymalizacja konsolidacji oparta na teorii informacji

Z perspektywy teorii informacji, proces konsolidacji można zoptymalizować poprzez maksymalizację zawartości informacyjnej przy jednoczesnej minimalizacji redundancji [158]:

$$\max_{C} I(C; D_{skonsolidowane}) - \lambda \cdot R(C)$$

gdzie:

- $I(C; D_{skonsolidowane})$ reprezentuje informację wzajemną między reprezentacją skonsolidowaną C a danymi oryginalnymi
- $R(C)$ oznacza miarę redundancji w danych skonsolidowanych
- λ jest parametrem kompromisu między kompletnością a zwięzłością

5.6. Zarządzanie tokenizacją i ekonomia LLM w aplikacjach bezpieczeństwa

Teoretyczne podstawy ekonomii tokenów w aplikacjach LLM

Praktyczne wykorzystanie LLM w aplikacjach przedsiębiorczych wymaga zaawansowanego zrozumienia ekonomii tokenów i jej wpływu na skalowalność analizy [159]. Problem można sformalizować jako optymalizację z ograniczeniami:

$$\max_{j,t} V(j) \quad \text{przy ograniczeniu} \quad C(t) \leq B, \quad t = f(j)$$

gdzie:

- $V(j)$ reprezentuje wartość analityczną jako funkcję jakości analizy
- $C(t)$ oznacza koszt jako funkcję zużycia tokenów
- B jest ograniczeniem budżetowym
- $t = f(j)$ reprezentuje relację między głębokością analizy a wymaganiami tokenowymi

Outlines jako struktura optymalizacji tokenowej

Biblioteka Outlines wprowadza fundamentalną zmianę w podejściu do ekonomii tokenowej w kontekście analizy bezpieczeństwa, oferując teoretyczne podstawy dla systematycznej optymalizacji procesów generowania treści przez LLM. Kluczową innowacją tej biblioteki jest możliwość przekształcenia tradycyjnie jednorazowych, kosztownych konstrukcji promptów w wielokrotnie wykorzystywane struktury analityczne.

Mechanizm ponownego wykorzystania szablonów stanowi fundament ekonomicznej efektywności systemu. W przeciwieństwie do tradycyjnych podejść, gdzie każda analiza wymaga konstrukcji dedykowanego prompta od podstaw, Outlines umożliwia amortyzację kosztów konstrukcji promptów w wielu przypadkach analizy. Szczególnie w środowiskach operacyjnych, gdzie podobne typy oceny bezpieczeństwa przeprowadzane są regularnie, oszczędności wynikające z wielokrotnego wykorzystania przemyślanie zaprojektowanych szablonów mogą być znaczące zarówno pod względem czasu, jak i kosztów tokenowych.

Strukturyzowane generowanie stanowi kolejny obszar optymalizacji, gdzie Outlines wymusza określone formaty wyjściowe bezpośrednio w procesie generowania, eliminując potrzebę kosztownego post-processingu. Tradycyjne podejścia często wymagają dodatkowych iteracji przetwarzania w celu uzyskania pożądanej struktury danych wyjściowych, co generuje dodatkowe koszty tokenowe oraz zwiększa ryzyko wprowadzenia błędów w procesie transformacji. Poprzez wbudowaną kontrolę formatowania, Outlines redukuje te koszty do minimum, jednocześnie gwarantując spójność i przewidywalność wyników.

Implementacja logiki warunkowej w ramach biblioteki umożliwia dynamiczną adaptację promptów opartą na dostępnym kontekście, eliminując niepotrzebną rozwlekłość charakterystyczną dla statycznych

szablonów. System może automatycznie dostosować poziom szczegółowości analizy w zależności od poziomu skomplikowania analizowanego przypadku, dostępnych danych wejściowych oraz specyficznych wymagań analitycznych. Taka adaptacyjność nie tylko optymalizuje wykorzystanie tokenów, ale również poprawia jakość merytoryczną generowanych analiz poprzez lepsze dopasowanie do specyfiki każdego przypadku.

Kompozycja hierarchiczna reprezentuje najbardziej zaawansowany aspekt optymalizacji tokenowej w Outlines, oferując modularną konstrukcję promptów z precyzyjną kontrolą nad alokacją tokenów na różnych poziomach abstrakcji analitycznej. Podejście to umożliwia tworzenie złożonych struktur analitycznych, gdzie poszczególne moduły mogą być niezależnie optymalizowane pod względem tokenowym, a następnie komponowane w kompleksowe ramy analityczne. Taka modularność jest szczególnie wartościowa w analizie systemów bezpieczeństwa, gdzie różne aspekty oceny mogą wymagać różnego poziomu szczegółowości i różnych metodologii analitycznych.

Strategiczna alokacja tokenów w analizie wielokryteriowej

W złożonej analizie bezpieczeństwa wymagającej wielu perspektyw analitycznych, strategiczna alokacja tokenów może być modelowana jako problem optymalizacji wielokryteriowej:

$$\max_{t_1,\ldots,t_k} \left[\sum_{i=1}^{k} w_i \cdot V_i(t_i) \right] \quad \text{przy ograniczeniu} \quad \sum_{i=1}^{k} t_i \leq T_{\text{całkowite}}$$

gdzie:

- t_i reprezentuje alokację tokenów dla wymiaru analitycznego i
- $V_i(t_i)$ oznacza funkcję wartości dla wymiaru i
- w_i są wagami ważności odzwierciedlającymi priorytety organizacyjne
- $T_{\text{całkowite}}$ jest całkowitym dostępnym budżetem tokenowym

Outlines umożliwia implementację tej strategii poprzez logikę szablonów warunkowych, która automatycznie dostosowuje złożoność promptów do dostępnego budżetu tokenowego przy zachowaniu priorytetów analitycznych.

5.7. Synteza teoretyczna i implikacje metodologiczne

Charakter analizy bezpieczeństwa wspomaganej przez LLM

Integracja LLM z automatyczną analizą infrastruktury prowadzi do zdolności analitycznych, które jakościowo przekraczają sumę poszczególnych komponentów. Tego rodzaju charakter może być teoretycznie opisany jako:

$$\mathcal{A}_{LLM+Auto} \not\subset \mathcal{A}_{LLM} \cup \mathcal{A}_{Auto}$$

gdzie $\mathcal{A}_{LLM+Auto}$ reprezentuje zdolności analityczne zintegrowanego systemu, a $\mathcal{A}_{LLM}$ i $\mathcal{A}_{Auto}$ oznaczają zdolności izolowanego LLM i automatycznej analizy odpowiednio.

Teoretyczne ograniczenia i zakres stosowalności

Mimo znaczącego potencjału, analiza bezpieczeństwa oparta na LLM napotyka na znaczne ograniczenia teoretyczne, które wymagają szczególnej uwagi w kontekście aplikacji o krytycznym znaczeniu dla bezpieczeństwa.

Podstawowe ograniczenie epistemologiczne wynika z faktu, że LLM mogą operować wyłącznie w ramach wzorców obecnych w danych treningowych. Oznacza to, że modele te nie są w stanie generować rzeczywiście nowej wiedzy wykraczającej poza istniejące paradygmaty analityczne. W kontekście dynamicznie ewoluującego środowiska zagrożeń, gdzie nowe typy ataków czy niekonwencjonalne scenariusze mogą wymagać innowacyjnego myślenia analitycznego, ograniczenie to staje się szczególnie istotne.

Wyzwania związane z interpretowalnością stanowią kolejny obszar problemowy. Charakter „czarnej skrzynki"rozumowania LLM utrudnia walidację procesów logicznych prowadzących do konkretnych wniosków analitycznych. W kontekstach bezpieczeństwa, gdzie konsekwencje błędnych decyzji mogą być katastrofalne, konieczność zrozumienia i weryfikacji ścieżek rozumowania modelu staje się kwestią krytyczną. Problem ten jest szczególnie istotny w środowiskach regulowanych, gdzie wymaga się pełnej audytowalności procesów decyzyjnych.

Ograniczenia techniczne związane z oknem kontekstu stanowią praktyczne wyzwanie w analizie kompleksowych scenariuszy bezpieczeństwa. Istotne ograniczenia w przetwarzaniu obszernego kontekstu mogą prowadzić do fragmentaryzacji analizy i utracenia holistycznego podejścia niezbędnego w kompleksowej ocenie systemów ochrony. Ta problematyka jest szczególnie widoczna przy analizie rozległych systemów infrastruktury krytycznej, gdzie liczba zmiennych i ich wzajemnych zależności może przekraczać możliwości skutecznego przetwarzania.

Wrażliwość na dryf dziedzinowy (domain drift) stanowi długoterminowe wyzwanie dla stabilności analitycznej. Degradacja wydajności modeli w sytuacjach, gdy kontekst operacyjny znacząco różni się od rozkładu treningowego, może prowadzić do stopniowej utraty adekwatności analiz bez wyraźnych sygnałów ostrzegawczych. Jest to szczególnie problematyczne w szybko ewoluujących środowiskach zagrożeń, gdzie zmienność taktyk przeciwnika może szybko dezaktualizować nawet najnowsze modele.

Struktura walidacji i zapewnienia jakości

Odpowiedź na przedstawione wyzwania wymaga opracowania kompleksowej struktury zapewnienia jakości, która może złagodzić ograniczenia techniczne poprzez systematyczne podejście metodologiczne.

Walidacja wieloperspektywiczna stanowi fundament takiego podejścia, wymagając systematycznego sprawdzania krzyżowego wniosków analitycznych względem niezależnych podejść metodologicznych. Proces ten powinien obejmować nie tylko porównanie z tradycyjnymi metodami analizy, ale również konfrontację z alternatywnymi modelami LLM oraz ekspercką oceną zewnętrzną. Taka triangulacja metodologiczna może znacząco zwiększyć wiarygodność generowanych analiz.

Kwantyfikacja niepewności stanowi niezbędny element transparentnego raportowania analitycznego. Eksplicytne modelowanie poziomów zaufania w generowanych rekomendacjach umożliwia podejmowa-

nie świadomych decyzji przez użytkowników końcowych oraz umożliwia systematyczne śledzenie obszarów o podwyższonym ryzyku analitycznym. Implementacja mechanizmów probabilistycznej oceny wiarygodności może znacząco poprawić użyteczność praktyczną generowanych analiz.

Mechanizmy pętli eksperckiej (ang. expert loop) wymagają systematycznej integracji ekspertyzy ludzkiej, szczególnie w krytycznych punktach decyzyjnych procesu analitycznego. Nie oznacza to prostego „zatwierdzania" przez eksperta wyników generowanych przez LLM, lecz przemyślaną strukturę współpracy, w której doświadczenie analityczne człowieka uzupełnia możliwości obliczeniowe modelu. Taka synergia może prowadzić do wykroczenia poza ograniczenia każdego z podejść stosowanego indywidualnie.

Walidacja czasowa stanowi długoterminowy element zapewnienia jakości, wymagający systematycznego śledzenia predykcji analitycznych względem rzeczywistych wyników bezpieczeństwa. Proces ten umożliwia nie tylko weryfikację trafności konkretnych analiz, ale również identyfikację systematycznej stronniczości (biasów) czy degradacji wydajności modelu w czasie. Implementacja takiej walidacji wymaga jednak opracowania standardowych metryk oraz systematycznych procesów gromadzenia danych zwrotnych z praktyki operacyjnej.

Podsumowanie teoretycznych podstaw LLM w kontekście bezpieczeństwa

Wykorzystanie LLM w automatyzacji oceny systemów bezpieczeństwa fizycznego łączy w sobie różne podejścia analityczne. Inżynieria promptów z wykorzystaniem szablonów oraz bibliotek do strukturyzacji danych (takich jak np. "outlines") tworzy podstawy dla systematycznej, powtarzalnej analizy. Jednocześnie zdolności LLM do rozpoznawania wzorców pozwalają odkrywać nieoczywiste problemy i słabości występujące w systemach bezpieczeństwa.

W niniejszej pracy w kontekście LLM zastosowano:

- **Jeden z wiodących modeli LLM** jako główny model do analizy semantycznej wykrytych elementów i interpretacji wyników grafowych
- **Bibliotekę "outlines"** do strukturyzowanej generacji raportów zgodnych z predefiniowanymi schematami JSON
- **Systematyczne szablony promptów** dla różnych etapów analizy (interpretacja planów, ocena ścieżek, generowanie rekomendacji)
- **Kontrolowaną generację** zapewniającą spójność formatowania i kompletność wyników analitycznych

Istotnym wydaje się nasza demonstracja, że zaawansowana konsolidacja danych połączona z ich strukturyzacją może transformować fragmentaryczną analizę techniczną w kompleksowe spostrzeżenia strategiczne. LLM pełni rolę semantycznego interpretatora, który otrzymuje strukturalne dane z wcześniejszych etapów procesu (wykryte elementy, ścieżki grafowe, parametry degradacji) i przekształca je w zrozumiałe analizy bezpieczeństwa z uwzględnieniem kontekstu domenowego.

Analiza teoretyczna wskazuje, że optymalne wdrożenie tego podejścia wymaga starannego planowania zarządzania tokenami kontekstowymi oraz strukturyzacji promptów, które mają na celu osiągnięcie równowagi między głębokością analityczną a ograniczeniami modelu. Szczególne znaczenie ma projektowanie szablonów promptów umożliwiających LLM efektywną interpretację wyników technicznych przy zachowaniu deterministycznej struktury wyjściowej poprzez kontrolowaną generację.

6. Implementacja narzędzia AI-SecPA

(AI – Security Paths Analyzer)

Na podstawie założeń teoretycznych i metodologii opracowanej w niniejszej rozprawie, dotyczącej oceny efektywności Cyber-Fizycznych Systemów Ochrony (CPPS) z uwzględnieniem degradacji w czasie, utworzono praktyczne narzędzie implementacyjne. Głównym celem tego narzędzia jest umożliwienie kompleksowej oceny skuteczności systemów bezpieczeństwa fizycznego z uwzględnieniem aspektów cyberbezpieczeństwa oraz procesów degradacyjnych zachodzących w czasie eksploatacji.

Narzędzie zostało skonstruowane w oparciu o założenia zarządzania ryzykiem, gdzie fundamentalnym podejściem jest porównanie ryzyka inherentnego (nieodłącznego) z ryzykiem rezydualnym. Implementacja ta pozwala na praktyczne zastosowanie teoretycznych podstaw przedstawionych w rozprawie, umożliwiając użytkownikom przeprowadzenie systematycznej analizy efektywności systemów ochrony.

Aplikacja została zaprojektowana z myślą o dostępności dla szerokiego grona użytkowników - może być uruchomiona lokalnie na komputerze pracownika, co eliminuje konieczność posiadania infrastruktury sieciowej. Choć wersja sieciowa nie stanowi problemu dla większych korporacji, dla większości organizacji wystarczający jest zakup komputera lub laptopa z obsługą GPU lub procesorów nowej generacji wspierających rozwiązania sztucznej inteligencji.

Narzędzie prowadzi użytkownika krok po kroku przez czytelny proces analizy, jednocześnie dostarczając konkretne wyniki i rezultaty każdego etapu oceny. Takie podejście zapewnia transparentność procesu i umożliwia użytkownikom zrozumienie mechanizmów wpływających na końcową ocenę efektywności systemu.

6.1. Część I: Przygotowanie danych i kontekstu

Ta część procesu obejmuje gromadzenie i przygotowanie wszystkich danych wejściowych niezbędnych do przeprowadzenia kompleksowej oceny efektywności systemów bezpieczeństwa. Etap przygotowania składa się z akwizycji dokumentacji technicznej obiektu oraz zebrania danych kontekstowych, które nie są widoczne na planach architektonicznych, a mają istotne znaczenie dla prawidłowej oceny.

Etap I: Pobranie rysunku podlegającego ocenie

Pierwszym krokiem w procesie oceny jest wczytanie dokumentacji graficznej obiektu podlegającego analizie. Na tym etapie system przyjmuje rysunek techniczny lub plan obiektu **bez uwzględnienia elementów systemów bezpieczeństwa**, co stanowi kluczowy element metodologii zarządzania ryzykiem.

Metodologia zarządzania ryzykiem - ryzyko inherentne

Podejście to wynika bezpośrednio ze standardowych procedur zarządzania ryzykiem [67, 160], gdzie:

I. Ryzyko inherentne (nieodłączne) - reprezentuje poziom ryzyka występujący w środowisku bez wdrożonych elementów kontrolujących skutki ryzyka. W kontekście systemów ochrony fizycznej oznacza to ocenę podatności obiektu na podstawie jego charakterystyki architektonicznej, lokalizacji i funkcji, przy założeniu braku jakichkolwiek systemów bezpieczeństwa.

II. Ryzyko rezydualne - stanowi poziom ryzyka pozostający po wdrożeniu środków kontrolnych i zabezpieczeń. W dalszych etapach analizy będą uwzględniane rzeczywiste systemy ochrony zainstalowane w obiekcie.

Takie podejście umożliwia:

- Obiektywną ocenę podstawowej podatności obiektu
- Porównanie skuteczności różnych wariantów systemów bezpieczeństwa
- Kwantyfikację rzeczywistej wartości dodanej przez wdrożone zabezpieczenia
- Optymalizację kosztów poprzez identyfikację najbardziej krytycznych elementów wymagających ochrony

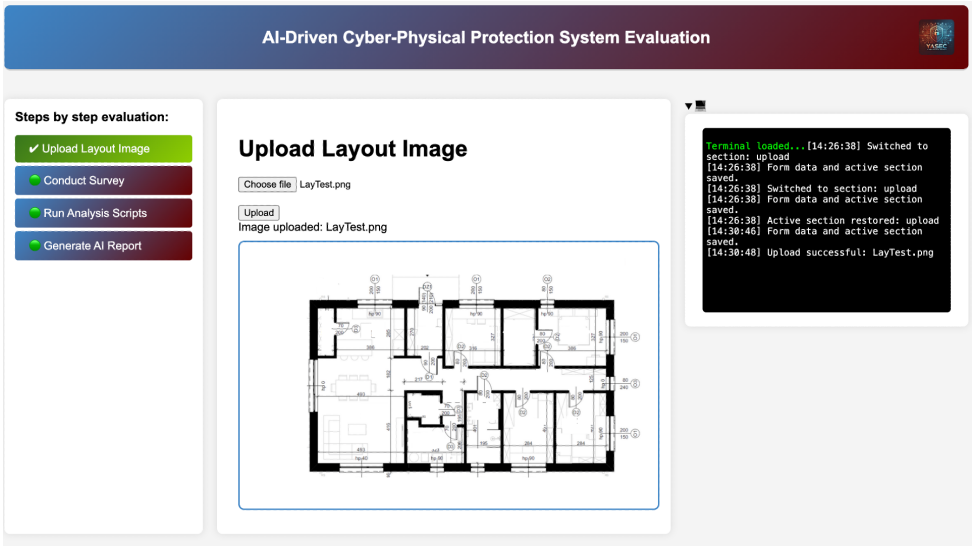
Analiza obrazu z wykorzystaniem algorytmów YOLO

Załadowany rysunek jest poddawany analizie z wykorzystaniem algorytmów detekcji obiektów YOLO (You Only Look Once), które zostały przystosowane do rozpoznawania elementów architektonicznych. System automatycznie identyfikuje i klasyfikuje:

- **Elementy konstrukcyjne**: ściany, słupy, belki
- **Otwory komunikacyjne**: drzwi, okna, bramy
- **Podziały funkcjonalne**: pomieszczenia, korytarze, klatki schodowe

Proces ten pozwala na utworzenie cyfrowego modelu obiektu, który stanowi podstawę do dalszej analizy ścieżek dostępu i identyfikacji potencjalnych punktów narażenia.

Etap akwizycji podkładu budowlanego jest przedstawiony na rysunku 6.1. Przedstawiony przykład stanowi plan komercyjnego obiektu podlegającego ochronie z wyraźnie zaznaczonymi pomieszczeniami i ich wymiarami. Rysunek nie zawiera żadnych elementów systemów bezpieczeństwa - jest to celowe podejście zgodne z metodologią oceny ryzyka inherentnego. Dzięki algorytmom YOLO system będzie w stanie automatycznie wykryć i sklasyfikować elementy architektoniczne, tworząc podstawę do dalszej analizy podatności obiektu.



Rysunek 6.1: Etap akwizycji podkładu budowlanego - interfejs wczytywania planu komercyjnego obiektu bez elementów systemów bezpieczeństwa zgodnie z metodologią oceny ryzyka inherentnego.

Ten pierwszy etap stanowi fundament dla całego procesu oceny, umożliwiając systemowi zrozumienie podstawowej struktury chronionego obiektu i zidentyfikowanie naturalnych barier oraz potencjalnych ścieżek dostępu, które będą analizowane w kolejnych krokach procedury ewaluacyjnej.

Etap II: Pobieranie danych kontekstowych - Security Survey

Drugi etap procesu oceny stanowi kluczowy element uzupełniający informacje pozyskane z analizy planów budynku. Etap ten koncentruje się na pobraniu danych kontekstowych, które nie są widoczne na rysunkach technicznych, a mają istotne znaczenie dla prawidłowej oceny efektywności systemów bezpieczeństwa. Bezpieczeństwo wymaga kontekstu - ten sam obiekt w różnych lokalizacjach, nawet z identycznymi zabezpieczeniami, będzie charakteryzował się odmiennym poziomem efektywności. Podobnie, ten sam layout przy użyciu różnych materiałów budowlanych daje różne wyniki efektywności, co wydaje się oczywiste - przejście przez ścianę z płyt gipsowo-kartonowych wymaga innego czasu ataku niż przejście przez ścianę z cegły.

Metodologia Security Survey

Ankiety bezpieczeństwa (ang. Security Survey) stanowią powszechnie stosowaną metodę w analizach bezpieczeństwa [161, 162]. Security Survey to systematyczny proces zbierania informacji o zasobach podlegających ochronie, istniejących zagrożeniach oraz aktualnych środkach bezpieczeństwa w celu identyfikacji słabych punktów i opracowania rekomendacji dotyczących poprawy poziomu ochrony. W kontekście niniejszego narzędzia, Security Survey służy do pozyskania trzech głównych kategorii danych kontekstowych:

Dane techniczne systemów bezpieczeństwa

Pierwsza kategoria obejmuje szczegółowe informacje techniczne dotyczące elementów systemów bezpieczeństwa, które nie są możliwe do odczytania z planów budynku:

- **Specyfikacje techniczne urządzeń**: parametry techniczne czujek (PIR, sejsmicznych, tłuczenia szkła, kontaktronów magnetycznych), kamer CCTV, systemów kontroli dostępu
- **Wiek i stan techniczny komponentów**: data instalacji, historia konserwacji, czas eksploatacji poszczególnych elementów
- **Klasy i certyfikaty bezpieczeństwa**: poziomy odporności zgodnie z normami branżowymi (np. EN 50131 dla systemów sygnalizacji włamania, EN 50133 dla systemów kontroli dostępu)
- **Parametry degradacji**: aktualne wskaźniki sprawności technicznej uwzględniane w modelu degradacji przedstawionym w rozprawie

Elementy ochrony osobowej

Druga kategoria dotyczy zasobów ludzkich zaangażowanych w proces ochrony:

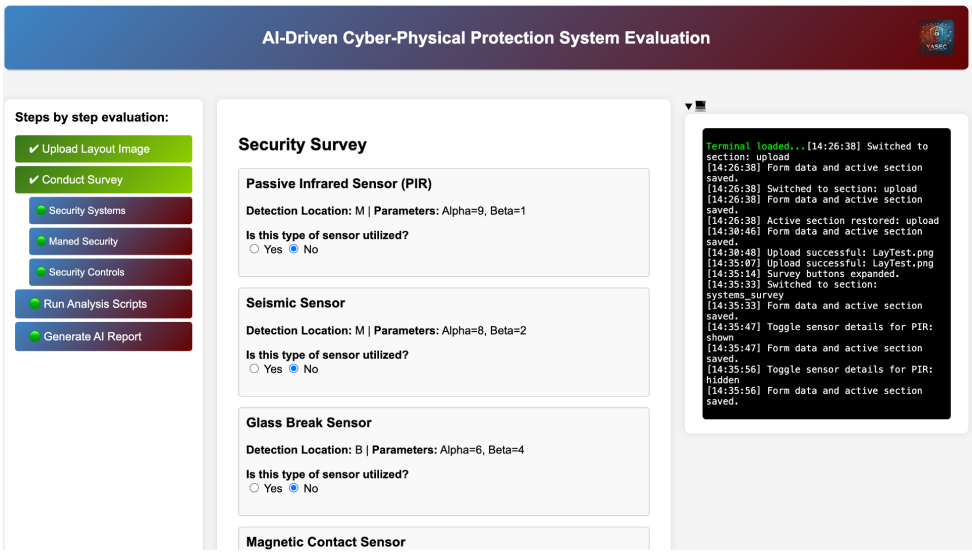
- **Posterunki ochrony fizycznej**: liczba stanowisk, rozmieszczenie, harmonogramy pracy
- **Grupy interwencyjne**: dostępność, czas reakcji, wyposażenie, procedury działania
- **Współpraca ze służbami zewnętrznymi**: czas dotarcia Policji, Straży Pożarnej, pogotowia ratunkowego
- **Kompetencje personelu**: poziom wyszkolenia, certyfikacje, doświadczenie operacyjne

Kontrole organizacyjne i kontekst geopolityczny

Trzecia kategoria obejmuje aspekty organizacyjne oraz szerszy kontekst środowiskowy:

- **Polityki i standardy bezpieczeństwa**: wewnętrzne regulacje organizacji, procedury bezpieczeństwa, tzw. baseline standards dla określonej branży
- **Kontekst geopolityczny**: aktualna sytuacja bezpieczeństwa w regionie pozyskiwana za pośrednictwem API do wybranych platform informacyjnych i ich archiwów
- **Charakterystyka otoczenia**: analiza lokalnych zagrożeń, statystyki przestępczości, specyfika regionu
- **Wymagania regulacyjne**: obowiązujące przepisy prawne, standardy branżowe, wymogi zgodności z prawem

Etap drugi procesu - pobieranie danych niewidocznych na planach budynku, a tworzących kontekst bezpieczeństwa - jest przedstawiony na rysunku 6.2. Interfejs umożliwia systematyczne wprowadzanie informacji o różnych typach czujek i systemów detekcji wraz z ich parametrami technicznymi. W prawej części ekranu widoczna jest konsola systemowa pokazująca przebieg procesu zbierania danych kontekstowych.



Rysunek 6.2: Etap drugi - pobieranie danych niewidocznych na planach budynku, a tworzących kontekst bezpieczeństwa. Interfejs Security Survey umożliwia wprowadzanie szczegółowych informacji technicznych o systemach detekcji.

Zebranie kompleksowych danych kontekstowych w ramach Security Survey stanowi niezbędny warunek dla prawidłowego funkcjonowania modelu degradacji oraz algorytmów oceny efektywności CPPS. Dane te są następnie integrowane z informacjami pozyskanymi z analizy planów budynku, tworząc holistyczny obraz systemu ochrony, który uwzględnia zarówno aspekty techniczne, organizacyjne, jak i środowiskowe wpływające na końcową ocenę bezpieczeństwa obiektu.

6.2. Część II: Detekcja i analiza obrazu

Ta część procesu wykorzystuje zaawansowane algorytmy komputerowej analizy obrazu i uczenia maszynowego do automatycznej analizy planów architektonicznych. Obejmuje detekcję elementów strukturalnych, ich weryfikację przez eksperta oraz przekształcenie w strukturyzowane dane przestrzenne gotowe do analizy grafowej.

Etap III: Detekcja elementów architektonicznych i post-processing

Po zakończeniu etapu Security Survey system przechodzi do fazy analizy obrazu i detekcji elementów architektonicznych z wykorzystaniem wcześniej załadowanego planu budynku. Ten etap obejmuje zastosowanie algorytmów YOLO do identyfikacji kluczowych komponentów struktury obiektu oraz zaawansowany post-processing mający na celu przygotowanie danych do dalszej analizy grafowej.

Proces detekcji z wykorzystaniem YOLO

Na tym etapie model wykrywa zdefiniowane klasy obiektów architektonicznych, które zostały wcześniej wyszczególnione w sekcji 3.2. Algorytm YOLO analizuje załadowany plan budynku i identyfikuje następujące elementy:

- **Ściany zewnętrzne i wewnętrzne**: struktury nośne i działowe
- **Otwory komunikacyjne**: drzwi różnych typów, okna, bramy
- **Elementy funkcjonalne**: pomieszczenia, korytarze, klatki schodowe

Post-processing - usuwanie duplikatów i segmentacja obiektów

Po zakończeniu procesu detekcji system dokonuje zaawansowanego post-processingu, dzięki któremu osiąga dwa kluczowe cele:

Eliminacja duplikatów

Algorytmy YOLO mogą generować wielokrotne detekcje tego samego obiektu z różnymi poziomami pewności. System implementuje algorytmy Non-Maximum Suppression (NMS) oraz dodatkowe heurystyki w celu usunięcia nakładających się detekcji i zachowania jedynie najbardziej wiarygodnych wyników.

Segmentacja obiektów dzielonych

Kluczowym elementem post-processingu jest inteligentna segmentacja wykrytych obiektów, szczególnie ścian, które są dzielone przez inne elementy architektoniczne. Ten proces jest niezbędny do prawidłowego utworzenia grafu reprezentującego potencjalne ścieżki ataku.

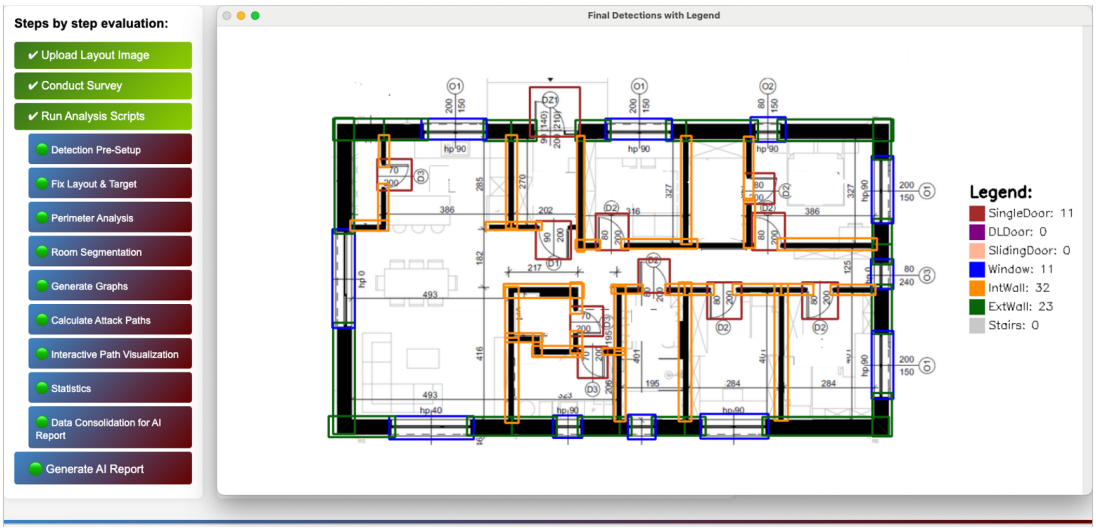
Problem segmentacji ścian: Gdy ściana zewnętrzna jest dzielona przez ścianę wewnętrzną, powstaje sytuacja wymagająca utworzenia dwóch węzłów grafu na ścianie zewnętrznej. Ma to istotne znaczenie dla późniejszej analizy, ponieważ:

- **Kontekst geograficzny ataku**: Po której stronie podziału rozpoczyna się atak ma znaczenie dla oceny prawdopodobieństwa detekcji
- **Różne poziomy zabezpieczeń**: Poszczególne segmenty mogą być chronione przez różne systemy detekcji
- **Wielopoziomowa analiza**: Umożliwia precyzyjną ocenę każdego segmentu struktury obronnej

Algorytm segmentacji: System implementuje algorytm geometryczny, który:

1. Identyfikuje punkty przecięcia między wykrytymi obiektami
2. Dzieli długie struktury (ściany, ogrodzenia) na logiczne segmenty
3. Tworzy unikalne identyfikatory dla każdego segmentu
4. Zachowuje informacje o powiązaniach między segmentami tego samego obiektu bazowego

Wynik procesu detekcji i post-processingu jest przedstawiony na rysunku 6.3. Plan budynku został poddany analizie, a wykryte elementy architektoniczne zostały oznaczone kolorami zgodnie z legendą. Widoczne są różne typy drzwi (SingleDoor, SlideDoor), okna (Window), ściany (Wall) oraz inne elementy strukturalne. Każdy wykryty obiekt otrzymał unikalny identyfikator umożliwiający dalszą analizę.



Rysunek 6.3: Wynik detekcji elementów architektonicznych z legendą. Poszczególne klasy obiektów zostały oznaczone różnymi kolorami: SingleDoor (czerwony), SlideDoor (fioletowy), Window (niebieski), Wall (zielony) oraz inne elementy według legendy.

Przygotowanie do analizy grafowej

Rezultatem tego etapu jest zestaw precyzyjnie zidentyfikowanych i posegmentowanych obiektów architektonicznych, z których każdy posiada:

- **Współrzędne geometryczne**: dokładne położenie na planie
- **Klasyfikację typu**: zgodną z predefiniowanymi kategoriami
- **Unikalny identyfikator**: umożliwiający śledzenie w dalszych etapach
- **Relacje sąsiedztwa**: informacje o połączeniach z innymi elementami
- **Parametry segmentacji**: dla obiektów podzielonych na mniejsze części

Te dane stanowią podstawę dla kolejnego etapu - tworzenia grafu reprezentującego wszystkie możliwe ścieżki dostępu do chronionego obiektu, który uwzględnia zarówno fizyczne bariery, jak i potencjalne punkty przełamania zabezpieczeń.

Etap IV: Manualna korekta wyników detekcji

Pomimo zaawansowanych algorytmów uczenia maszynowego i wysokiej skuteczności modeli YOLO, proces automatycznej detekcji elementów architektonicznych nie jest w pełni bezbłędny. Ponadto, nawet w przypadku perfekcyjnej detekcji, nie ma gwarancji, że rysunek odpowiada w 100% rzeczywistemu stanowi fizycznego budynku, do którego się odnosi. Z tego powodu niezbędnym było zaimplementowanie narzędzia umożliwiającego człowiekowi ręczną poprawę potencjalnych błędów i nieścisłości.

Filozofia wspomagania eksperta

Narzędzie zostało zaprojektowane z założeniem wspomagania eksperta, a nie jego zastąpienia. Choć istnieje element manualnej korekty wyników zwracanych przez model, efekt związany z nakładem na poprawę w stosunku do tworzenia kompletnego drzewa ścieżek dostępu manualnie jest nieporównywalnie bardziej efektywny. Takie podejście jest zgodne z najlepszymi praktykami implementacji systemów AI w zastosowaniach krytycznych, gdzie ostateczna decyzja i weryfikacja pozostaje w gestii eksperta [163, 164].

Konieczność korekty manualnej

Potrzeba manualnej interwencji wynika z kilku fundamentalnych ograniczeń automatycznej detekcji:

Ograniczenia techniczne modeli detekcji

Algorytmy YOLO, mimo wysokiej skuteczności, napotykają trudności w specyficznych sytuacjach:

– Obiekty o niestandardowych kształtach lub proporcjach

– Elementy o słabym kontraście wizualnym na planach

– Nakładające się lub częściowo zasłonięte struktury

– Błędna klasyfikacja podobnych obiektów architektonicznych

Rozbieżności między dokumentacją a rzeczywistością

Plany architektoniczne nie zawsze odzwierciedlają aktualny stan obiektu:

– Modyfikacje wprowadzone po zakończeniu budowy

– Tymczasowe zmiany w strukturze budynku

– Elementy bezpieczeństwa dodane post-factum

– Różnice między projektem a wykonaniem

Specyfika kontekstu bezpieczeństwa

Ekspert posiada wiedzę niedostępną algorytmom:

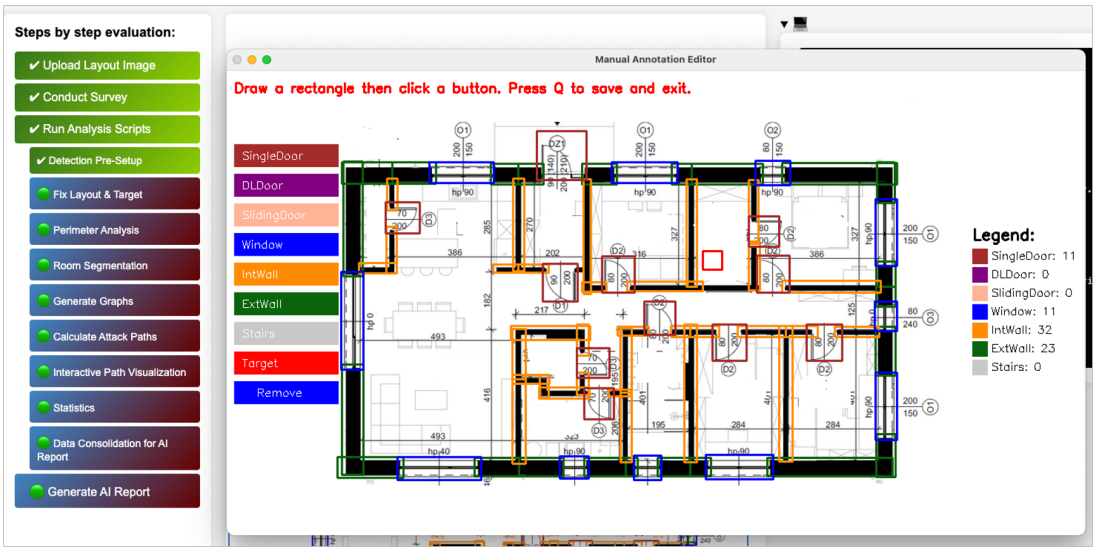
– Ukryte wejścia i przejścia techniczne

– Elementy o znaczeniu taktycznym

– Punkty szczególnie wrażliwe z perspektywy ochrony

– Lokalną specyfikę zagrożeń

Etap manualnej korekty wyników detekcji przedstawiono na rysunku 6.4. Interfejs umożliwia ekspertowi weryfikację i poprawę automatycznie wykrytych elementów architektonicznych przed przejściem do etapu analizy grafowej.



Rysunek 6.4: Etap manualnej korekty wyników detekcji - interfejs umożliwiający ekspertowi poprawę błędów automatycznej detekcji i dostosowanie wyników do rzeczywistego stanu obiektu.

Wartość dodana procesu korekty

Wprowadzenie etapu manualnej korekty przynosi wymierne korzyści:

Poprawa jakości danych wejściowych: Eliminacja błędów detekcji zapobiega propagacji nieprawidłowości do kolejnych etapów analizy, co bezpośrednio przekłada się na wiarygodność końcowych wyników.

Uwzględnienie wiedzy eksperckiej: Możliwość wprowadzenia informacji niedostępnych algorytmom, szczególnie dotyczących specyfiki bezpieczeństwa danego obiektu.

Zachowanie efektywności: Mimo elementu manualnego, całkowity nakład pracy pozostaje znacząco niższy niż w przypadku kompletnie ręcznej analizy, gdyż automatyczna detekcja dostarcza solidnej podstawy, wymagającej jedynie punktowych korekt.

Zwiększenie zaufania do wyników: Świadomość, że dane zostały zweryfikowane przez eksperta, zwiększa akceptowalność systemu w środowisku profesjonalnym.

Skorygowane i zweryfikowane dane stanowią podstawę dla kolejnego kluczowego etapu - generowania grafu reprezentującego wszystkie możliwe ścieżki dostępu do chronionego obiektu, gdzie jakość danych wejściowych ma bezpośredni wpływ na precyzję analizy podatności i skuteczność oceny systemów ochrony.

6.3. Część III: Modelowanie grafowe i analiza ścieżek ataku

W tej części wykryte elementy architektoniczne są przekształcane w graf topologiczny reprezentujący możliwe ścieżki przemieszczania się przeciwnika. System generuje względnie racjonalne scenariusze ataku, przeprowadza ich wizualizację oraz analizę statystyczną w celu identyfikacji krytycznych punktów infrastruktury.

Etap V: Wykrywanie obiektów obwodowych (perymetrowych)

Po zakończeniu manualnej korekty wyników detekcji system przechodzi do kolejnego kluczowego etapu - identyfikacji obiektów tworzących perymetr budynku. W tym kroku dokonuje się zaawansowanego post-processingu, który na poprawionym rysunku określa, które obiekty są elementami tworzącymi tak zwany perymetr budynku, ponieważ to one będą stanowić węzły grafu reprezentujące potencjalne punkty początkowe ścieżek ataku.

Koncepcja perymetru w analizie bezpieczeństwa

Perymetr stanowi pierwszą linię obrony obiektu chronionego i definiuje granicę między środowiskiem kontrolowanym a niekontrolowanym [162, 165]. W kontekście analizy ścieżek ataku, elementy perymetrowe pełnią podwójną funkcję:

- **Punkty inicjacji ataku**: reprezentują miejsca, w których przeciwnik może rozpocząć próbę naruszenia zabezpieczeń
- **Mechanizm filtrowania ścieżek**: umożliwiają ograniczenie teoretycznych ścieżek dostępu do tych, które można nazwać względnie racjonalnymi z perspektywy rzeczywistego scenariusza ataku

Algorytm identyfikacji perymetru

Proces wykrywania obiektów perymetrowych wykorzystuje analizę geometryczną skorygowanych wyników detekcji. System implementuje algorytm, który:

Klasyfikuje elementy zewnętrzne

Identyfikuje obiekty stanowiące fizyczną granicę budynku:

- Ściany zewnętrzne (ExWall) - główne bariery strukturalne
- Okna (Window) - potencjalne punkty dostępu o różnym poziomie zabezpieczenia
- Drzwi zewnętrzne (SingleDoor, SlideDoor) - kontrolowane punkty wejścia
- Inne otwory - elementy techniczne, wentylacyjne itp.

Analizuje graficzną ciągłość perymetru

Sprawdza spójność geometryczną i identyfikuje:

- Luki w perymetrze wymagające szczególnej uwagi
- Punkty styku różnych typów barier
- Obszary o potencjalnie obniżonej odporności

Przypisuje priorytety taktyczne

Ocenia względną atrakcyjność poszczególnych punktów z perspektywy przeciwnika:

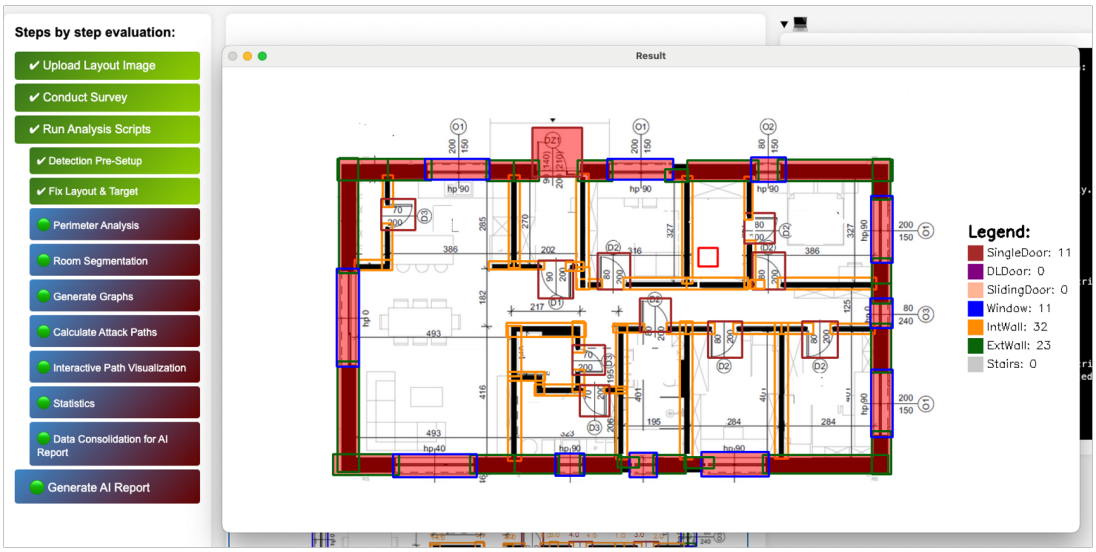
- Dostępność z zewnątrz

– Poziom widoczności i kontroli

– Bliskość celów o wysokiej wartości

– Możliwości dalszego przemieszczania się

Wynik procesu wykrywania obiektów perymetrowych jest przedstawiony na rysunku 6.5. Elementy tworzące perymetr budynku zostały wyróżnione czerwoną ramką, co pozwala na wizualną weryfikację poprawności automatycznej identyfikacji granic obiektu chronionego.



Rysunek 6.5: Wynik wykrywania obiektów perymetrowych - elementy tworzące zewnętrzną granicę budynku zostały oznaczone czerwoną ramką, definiując potencjalne punkty inicjacji ścieżek ataku.

Znaczenie dla analizy ścieżek ataku

Precyzyjna identyfikacja perymetru ma istotne znaczenie dla kolejnych etapów analizy:

Ograniczenie przestrzeni problemowej

Definicja perymetru pozwala na:

– Zawężenie analizy do realistycznych scenariuszy ataku

– Eliminację teoretycznych ścieżek niewykonalnych w praktyce

– Fokus na najbardziej prawdopodobnych kierunkach zagrożenia

Optymalizacja obliczeń grafowych

Ograniczenie punktów startowych:

– Redukuje złożoność obliczeniową algorytmów przeszukiwania ścieżek

– Umożliwia głębszą analizę rzeczywistych zagrożeń

– Przyspiesza proces generowania i oceny ścieżek

Podstawa dla analizy ryzyka

Zidentyfikowane elementy perymetrowe stanowią:

– Punkty odniesienia dla oceny efektywności systemów detekcji

– Podstawę do przypisania wartości prawdopodobieństwa inicjacji ataku

– Metodologia dla integracji z danymi Security Survey dotyczącymi konkretnych zabezpieczeń

Integracja z modelem degradacji

Obiekty perymetrowe, jako pierwsza linia obrony, są szczególnie narażone na procesy degradacji opisane w teoretycznych podstawach niniejszej rozprawy. System uwzględnia:

– Wiek i stan techniczny barier fizycznych

– Poziom zabezpieczenia elektronicznego poszczególnych segmentów

– Wpływ czynników środowiskowych na odporność perymetru

– Efektywność systemów detekcji w różnych warunkach eksploatacyjnych

Zidentyfikowane i sklasyfikowane obiekty perymetrowe stanowią kluczowe dane wejściowe dla kolejnego etapu - generowania kompletnego grafu ścieżek dostępu, który uwzględni wszystkie możliwe kombinacje ataków na zdefiniowany perymetr ochrony.

Etap VI: Segmentacja pomieszczeń i weryfikacja graficznej ciągłości

Na tym etapie dokonywana jest detekcja pomieszczeń layoutu, co również jest konieczne do stworzenia grafu dla względnie racjonalnych ścieżek ataku. Jest to również moment, w którym testowane jest, czy manualna poprawa przyniosła skutek, czy wymaga dopracowania. Dzięki temu ujawnia się to, co zostało pokazane przy pomocy czerwonych okręgów na rysunku - miejsca, gdzie nie ma graficznej ciągłości, a obiekty, których długość zostaje zwiększona właśnie dla jej zachowania, mają inny wymiar (dłuższy) niż przewidywany.

Znaczenie segmentacji pomieszczeń

Prawidłowa identyfikacja pomieszczeń stanowi kluczowy element analizy ścieżek ataku z kilku powodów:

Logika przemieszczania się przeciwnika

Pomieszczenia definiują naturalne jednostki przestrzenne, przez które przemieszcza się potencjalny intruz:

– Każde pomieszczenie stanowi węzeł w grafie możliwych ścieżek

– Połączenia między pomieszczeniami określają możliwe kierunki ruchu

– Wielkość i funkcja pomieszczenia wpływa na prawdopodobieństwo detekcji

Kontekst systemów bezpieczeństwa

Rozmieszczenie systemów detekcji często odpowiada strukturze pomieszczeń:

– Czujki PIR są zazwyczaj przypisane do konkretnych pomieszczeń

– Systemy kontroli dostępu działają na poziomie przejść między pomieszczeniami

– Kamery CCTV pokrywają zdefiniowane obszary funkcjonalne

Wykrywanie problemów graficznej ciągłości

Proces segmentacji pomieszczeń służy również jako mechanizm kontroli jakości poprzednich etapów. Algorytm wykrywa problemy w graficznej ciągłości struktury, które mogą prowadzić do błędnej oceny:

Identyfikacja luk strukturalnych

System automatycznie wykrywa miejsca, gdzie:

– Brakuje elementów łączących (np. fragmenty ścian)

– Występują nieuzasadnione przerwy w barierach

– Geometria obiektów nie tworzy spójnych granic pomieszczeń

Kompensacja przez wydłużanie obiektów

W miejscach wykrytych luk system podejmuje próbę automatycznej kompensacji poprzez:

– Wydłużanie istniejących elementów strukturalnych

– Interpolację brakujących fragmentów barier

– Tworzenie sztucznych połączeń dla zachowania topologii

Jednak przez tego rodzaju błąd tworzenie grafu i jego ocena może prowadzić do błędnej oceny całego systemu ochrony. Wynik segmentacji pomieszczeń i wykrytych problemów w graficznej ciągłości przedstawiono na rysunku 6.6. W lewej części widoczny jest plan z wykrytymi pomieszczeniami (oznaczonymi kolorami) oraz węzłami grafu. Czerwone okręgi wskazują miejsca problemów graficznej ciągłości, gdzie automatyczne wydłużenie elementów strukturalnych prowadzi do nieprawidłowej geometrii. Prawa część pokazuje tabelę z wykrytymi pomieszczeniami wraz z ich parametrami geometrycznymi.

Konsekwencje problemów z graficzną ciągłością

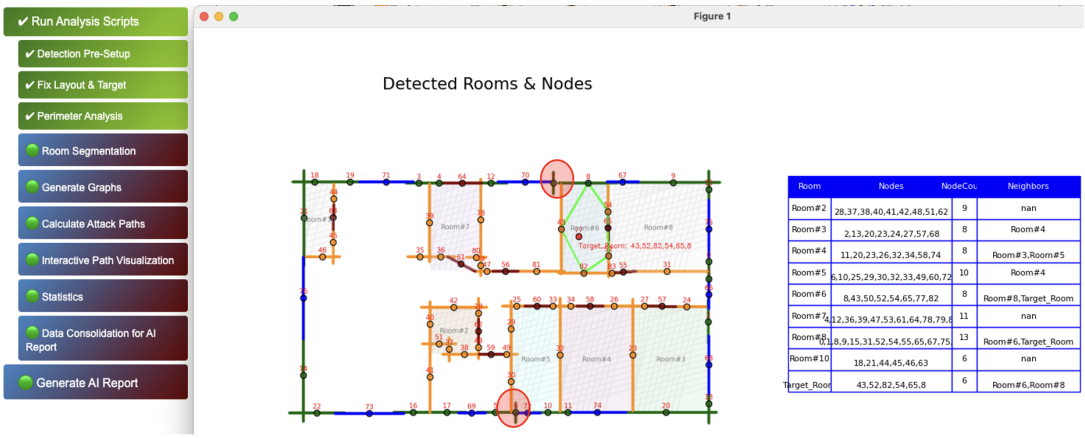
Niewykryte lub niepoprawione problemy z graficzną ciągłością mogą prowadzić do szeregu błędów w dalszej analizie:

Błędne ścieżki ataku

– Tworzenie teoretycznych przejść, które nie istnieją w rzeczywistości

– Pominięcie rzeczywistych barier na drodze przeciwnika

– Nieprawidłowa ocena czasów przejścia między pomieszczeniami



Rysunek 6.6: Segmentacja pomieszczeń z wykrytymi problemami w ich graficznej ciągłości - czerwone okręgi wskazują miejsca wymagające korekty, gdzie automatyczne wydłużenie elementów prowadzi do nieprawidłowej geometrii.

Błędna ocena systemów detekcji

- Niewłaściwe przypisanie czujek do pomieszczeń
- Błędne założenia dotyczące pokrycia systemów bezpieczeństwa
- Nieprawidłowa kalkulacja prawdopodobieństw detekcji

Propagacja błędów

- Problemy geometryczne wpływają na wszystkie kolejne etapy analizy
- Niedokładności kumulują się w końcowej ocenie efektywności
- Wyniki mogą prowadzić do błędnych decyzji inwestycyjnych

Konieczność iteracyjnej korekty

Na przykładzie widać, że rysunek należy poprawić jeszcze raz w dwóch miejscach, by wszystkie pomieszczenia zostały wykryte i poprawnie zmapowane. Ten etap często wymaga powrotu do fazy manualnej korekty w celu:

- Uzupełnienia brakujących elementów strukturalnych
- Korekty geometrii nieprawidłowo wydłużonych obiektów
- Dodania pominiętych barier lub otworów
- Weryfikacji poprawności segmentacji wszystkich pomieszczeń

Proces ten podkreśla iteracyjny charakter systemu i konieczność zachowania kontroli eksperta nad jakością danych na każdym etapie analizy. Tylko prawidłowo zidentyfikowane i zmapowane pomieszczenia mogą stanowić podstawę dla wiarygodnej analizy ścieżek ataku i oceny efektywności systemów bezpieczeństwa.

Etap VII: Generowanie grafu względnie racjonalnych ścieżek ataku

Po zakończeniu segmentacji pomieszczeń i weryfikacji graficznej ciągłości struktury system przechodzi do kluczowego etapu - tworzenia grafu reprezentującego względnie racjonalne ścieżki ataku. Na tym etapie dokonywana jest pierwsza istotna redukcja teoretycznie wszystkich ścieżek do ścieżek, które są względnie racjonalne - takich, które nie wychodzą z obiektu, a prowadzą od wejścia do celu ataku.

Koncepcja racjonalności ścieżek ataku

Racjonalna ścieżka ataku charakteryzuje się następującymi cechami:

- **Logiczny kierunek przemieszczania**: od punktu wejścia do celu bez zbędnych objazdów
- **Minimalizacja ekspozycji**: unikanie niepotrzebnego narażenia na detekcję
- **Efektywność czasowa**: preferencja dla krótszych ścieżek przy podobnym poziomie ryzyka
- **Wykonywalność fizyczna**: uwzględnienie rzeczywistych ograniczeń przestrzennych

Struktura grafu ataku

Graf kombinowany przedstawiony na rysunku 6.7 ilustruje trzy kategorie węzłów w strukturze ścieżek ataku:

Węzły perymetrowe (czerwone)

Reprezentują punkty wejścia do obiektu chronionego:

- Zlokalizowane na granicy perymetru bezpieczeństwa
- Stanowią potencjalne punkty inicjacji ataku
- Każdy węzeł perymetrowy może być punktem startowym niezależnej ścieżki

Węzły transferowe (żółte i pomarańczowe)

Znajdują się między punktem wejścia a fizycznym celem ataku:

- Reprezentują pomieszczenia, korytarze i inne przestrzenie komunikacyjne
- Umożliwiają transfer przeciwnika przez strukturę obiektu
- Stanowią punkty decyzyjne w wyborze dalszej drogi

Węzły docelowe (target)

Reprezentują cele ataku:

- Zasoby o wysokiej wartości wymagające ochrony
- Systemy krytyczne dla funkcjonowania organizacji
- Punkty końcowe wszystkich analizowanych ścieżek

Algorytm redukcji ścieżek

System implementuje zaawansowany algorytm filtrowania, który eliminuje nieracjonalne ścieżki na podstawie następujących kryteriów:

Kryterium kierunkowości

- Eliminacja ścieżek prowadzących z powrotem do punktu wejścia
- Usunięcie cykli nieprowadzących do celu
- Preferowanie ścieżek utrzymujących konsekwentny kierunek do celu

Kryterium efektywności

- Analiza stosunku długości ścieżki do postępu w kierunku celu
- Eliminacja nadmiernie wydłużonych tras alternatywnych
- Zachowanie różnorodności ścieżek przy podobnej efektywności

Kryterium wykonalności

- Weryfikacja fizycznej możliwości przejścia
- Uwzględnienie ograniczeń czasowych i logistycznych
- Analiza praktycznych aspektów przemieszczania się

Rysunek 6.7 przedstawia wizualizację wygenerowanego grafu względnie racjonalnych ścieżek ataku. Wizualizacja tworzona jest na potrzeby raczej dydaktyczno-poglądowe niż ze względu na faktyczną potrzebę dla procesu. Rysunek dobrze obrazuje węzły stanowiące punkty wejścia (czerwone), węzły znajdujące się w grafie między punktem wejścia a celem (żółte i pomarańczowe) oraz sam cel (czarny węzeł).

Znaczenie redukcji dla analizy

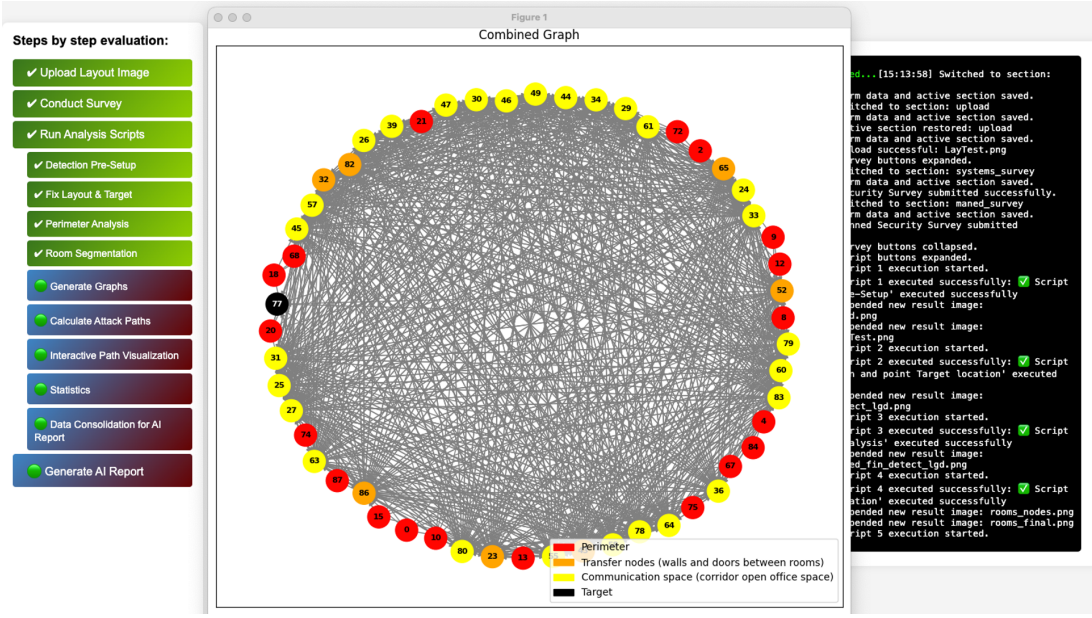
Proces redukcji teoretycznych ścieżek do racjonalnych ma kluczowe znaczenie dla dalszej analizy:

Optymalizacja obliczeniowa

- Znaczące zmniejszenie liczby ścieżek wymagających szczegółowej analizy
- Redukcja złożoności obliczeniowej algorytmów oceny ryzyka
- Umożliwienie głębszej analizy pozostałych, istotnych ścieżek

Zwiększenie trafności wyników

- Fokus na scenariuszach o rzeczywistym prawdopodobieństwie wystąpienia
- Eliminacja teoretycznych przypadków skrajnych
- Poprawa praktycznej użyteczności końcowych rekomendacji



Rysunek 6.7: Graf kombinowany przedstawiający względnie racjonalne ścieżki ataku - węzły perymetrowe (czerwone) stanowią punkty wejścia, węzły transferowe (żółte/pomarańczowe) reprezentują przestrzenie komunikacyjne, a gęsta sieć połączeń ilustruje wszystkie możliwe racjonalne ścieżki między wejściem a celem.

Podstawa dla oceny systemów bezpieczeństwa

– Identyfikacja kluczowych punktów wymagających zabezpieczenia

– Mapowanie systemów detekcji na realne ścieżki zagrożeń

– Umożliwienie precyzyjnej kalkulacji efektywności CPPS

Wygenerowany graf stanowi fundament dla kolejnych etapów analizy, w których każda z racjonalnych ścieżek zostanie poddana szczegółowej ocenie z uwzględnieniem modelu degradacji systemów bezpieczeństwa oraz danych kontekstowych zebranych w ramach Security Survey.

Etap VIII: Interaktywna wizualizacja ścieżek ataku

Po wygenerowaniu grafu względnie racjonalnych ścieżek ataku, system przechodzi do etapu interaktywnej wizualizacji, który umożliwia ekspertowi szczegółową analizę poszczególnych ścieżek zgodnie z ich rankingiem opłacalności. Wizualizacja ścieżek ataków zgodnie z ich kosztem stanowi narzędzie, które pozwala ekspertowi weryfikować dane wynikające z oceny oraz zrozumieć logikę działania systemu oceny.

Ranking opłacalności ścieżek

System automatycznie oblicza koszt każdej ścieżki ataku, uwzględniając:

– **Czas przemieszczania**: długość fizycznej drogi i trudność pokonania barier

– **Prawdopodobieństwo detekcji**: skumulowane ryzyko wykrycia na poszczególnych etapach

– **Efektywność systemów bezpieczeństwa**: uwzględnienie degradacji zgodnie z modelem przedstawionym w rozprawie

– **Kontekst operacyjny**: dane z Security Survey dotyczące konkretnych zabezpieczeń

W omawianym przypadku łącznie do celu doszło 115 ścieżek, które zostały uporządkowane według opłacalności z perspektywy potencjalnego przeciwnika.

Analiza porównawcza ścieżek

Rysunki 6.8 i 6.9 przedstawiają wizualizację dwóch skrajnych przypadków:

Ścieżka najbardziej opłacalna (Path 1: Cost 271.42)

Ścieżka oznaczona jako najkorzystniejsza dla przeciwnika charakteryzuje się:

– Niską złożonością topologiczną, z małą liczbą punktów przejścia

– Przejściem przez pomieszczenia o prawdopodobnie niższym poziomie monitoringu

– Wykorzystaniem naturalnych korytarzy komunikacyjnych

– Krótkim czasem ekspozycji na systemy detekcji

Ścieżka o niskiej opłacalności (Path 110: Cost 946.09)

Ścieżka znajdująca się blisko końca rankingu opłacalności wykazuje:

– Znacznie wydłużoną trasę z licznymi zmianami kierunku

– Przejście przez pomieszczenia o wysokim poziomie zabezpieczeń

– Większą ekspozycję na systemy detekcji

– Wyższe prawdopodobieństwo przerwania ataku

Funkcjonalność walidacyjna

Interaktywna wizualizacja pełni kluczową rolę w procesie walidacji wyników:

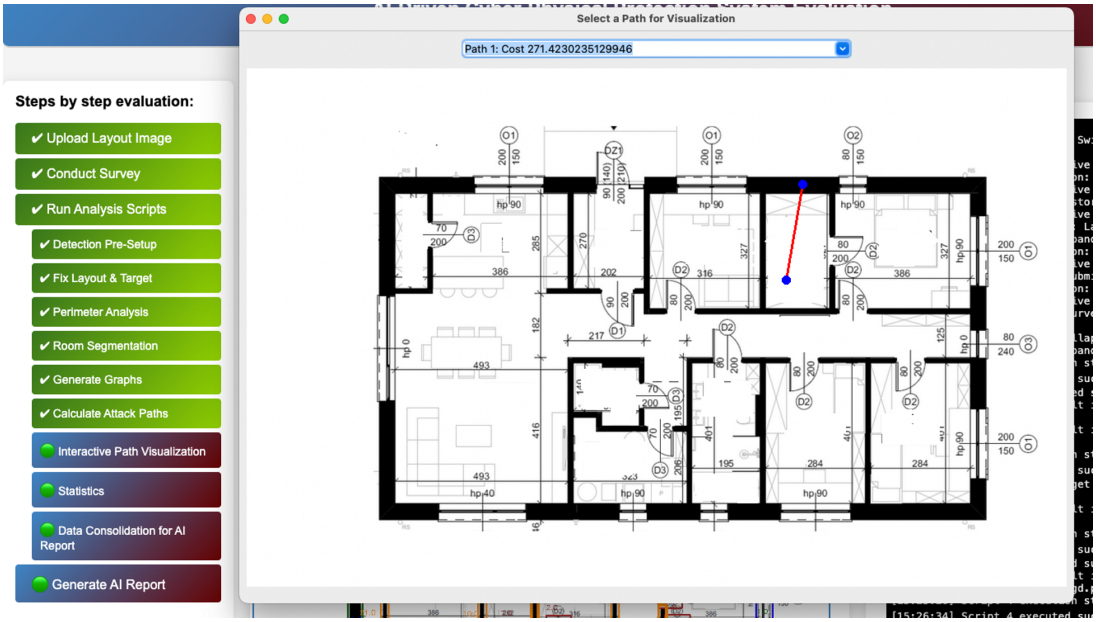
Weryfikacja logiki systemu

Ekspert może:

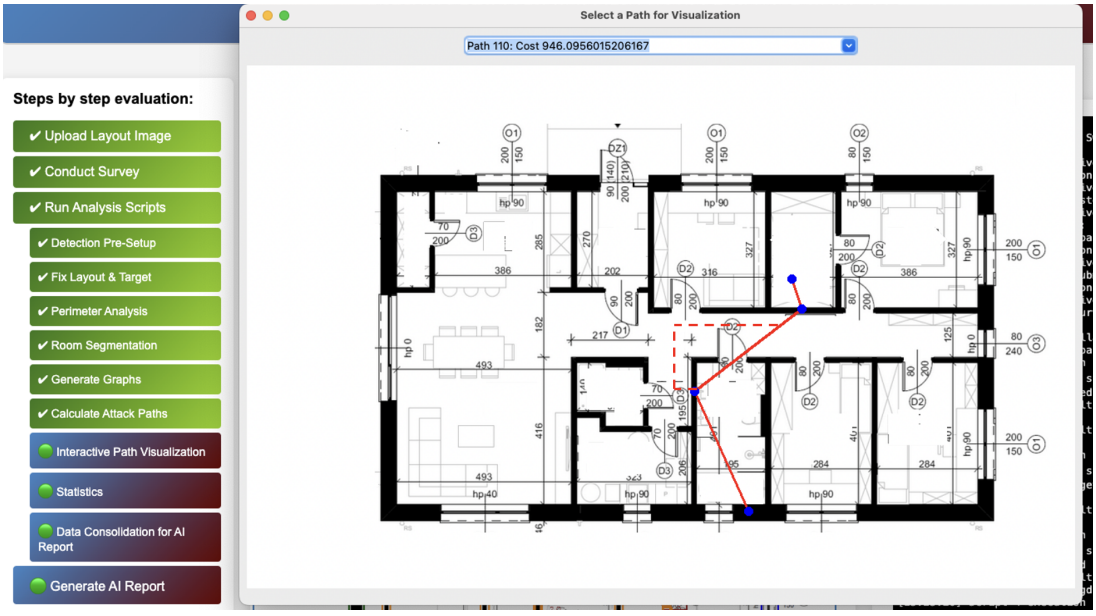
– Sprawdzić, czy najkorzystniejsze ścieżki rzeczywiście odpowiadają intuicji bezpieczeństwa

– Zidentyfikować potencjalne błędy w algorytmach oceny

– Walidować poprawność uwzględnienia danych kontekstowych



Rysunek 6.8: Wizualizacja najbardziej opłacalnej ścieżki ataku (Path 1: Cost 271.42) - linia ciągła pokazuje optymalną trasę z perspektywy przeciwnika, charakteryzującą się minimalnym kosztem i ryzykiem detekcji.



Rysunek 6.9: Wizualizacja ścieżki o niskiej opłacalności (Path 110: Cost 946.10) - linia przerywana ilustruje złożoną trasę o wysokim koszcie, wymagającą przejścia przez liczne bariery i systemy bezpieczeństwa.

Analiza wrażliwości

Porównanie skrajnych ścieżek pozwala na:

– Ocenę rozpięcia kosztów między najłatwiejszymi a najtrudniejszymi atakami

– Identyfikację krytycznych punktów wpływających na ranking

– Zrozumienie względnego znaczenia różnych czynników bezpieczeństwa

Optymalizacja systemów ochrony

Wizualizacja umożliwia:

– Identyfikację ścieżek wymagających wzmocnienia zabezpieczeń

– Priorytetyzację inwestycji w systemy bezpieczeństwa

– Ocenę efektywności istniejących rozwiązań

Interpretacja wyników

Linia przerywana na rysunkach pokazuje, jak czytać poszczególne fragmenty ścieżki ataku, ilustrując:

– Kolejność przejść między pomieszczeniami

– Punkty podejmowania decyzji o dalszej trasie

– Miejsca największego ryzyka detekcji

– Alternatywne opcje dostępne dla przeciwnika

Znaczna różnica w kosztach między ścieżką optymalną (271.42) a ścieżką o niskiej opłacalności (946.10) potwierdza skuteczność algorytmu różnicowania tras pod względem atrakcyjności dla potencjalnego przeciwnika. Ta rozpiętość wskazuje również na obecność zarówno słabych punktów w systemie ochrony, jak i obszarów o wysokiej skuteczności zabezpieczeń.

Interaktywna wizualizacja stanowi zatem nie tylko narzędzie analityczne, ale również instrument edukacyjny, umożliwiający ekspertom bezpieczeństwa głębsze zrozumienie dynamiki zagrożeń i mechanizmów ochrony w analizowanym obiekcie.

Etap IX: Analiza statystyk i identyfikacja krytycznych węzłów

Etap analizy statystyk umożliwia ekspertowi ocenę intensywności "ruchu"występującego na określonych węzłach i krawędziach ocenionych ścieżek ataku. Na tym etapie system przeprowadza szczegółową analizę częstotliwości wykorzystania poszczególnych elementów infrastruktury w najbardziej efektywnych scenariuszach ataku, co pozwala na identyfikację kluczowych punktów wymagających szczególnej uwagi z perspektywy bezpieczeństwa.

Metodologia analizy statystycznej

System agreguje dane z wszystkich wygenerowanych ścieżek, obliczając dla każdego węzła i krawędzi grafu:

Metryki częstotliwości wykorzystania

- **Częstotliwość przejść przez węzeł**: liczba ścieżek wykorzystujących dany punkt
- **Intensywność ruchu na krawędziach**: częstość wykorzystania połączeń między pomieszczeniami
- **Pozycja w rankingu**: udział w ścieżkach o różnych poziomach opłacalności
- **Krytyczność węzła**: wpływ na dostępność alternatywnych tras

Analiza koncentracji zagrożeń

- Identyfikacja tzw. "wąskich gardeł"w strukturze obiektu
- Wykrycie węzłów o kluczowym znaczeniu taktycznym
- Mapowanie naturalnych korytarzy infiltracji

Wizualizacja statystyk ruchu

Rysunek 6.10 przedstawia wizualizację pierwszych 15 ścieżek ataków ze wskazaniem krawędzi i węzłów w nich występujących. Graficzna reprezentacja wykorzystuje układ kołowy, który pozwala na:

Identyfikację wzorców ruchu

- **Węzły perymetrowe (czerwone)**: punkty wejścia najczęściej wykorzystywane przez efektywne ścieżki
- **Węzły transferowe (pomarańczowe)**: pomieszczenia o wysokiej częstotliwości przejść
- **Przestrzenie komunikacyjne (żółte)**: korytarze i obszary tranzytowe
- **Cel ('Target' - czarny węzeł)**: fizyczna lokalizacja celu wszystkich analizowanych ścieżek

Analiza intensywności połączeń

Grubość i kolor krawędzi reprezentują:

- Częstotliwość wykorzystania danego połączenia
- Względne znaczenie dla najefektywniejszych ścieżek
- Krytyczność dla ogólnej dostępności celu

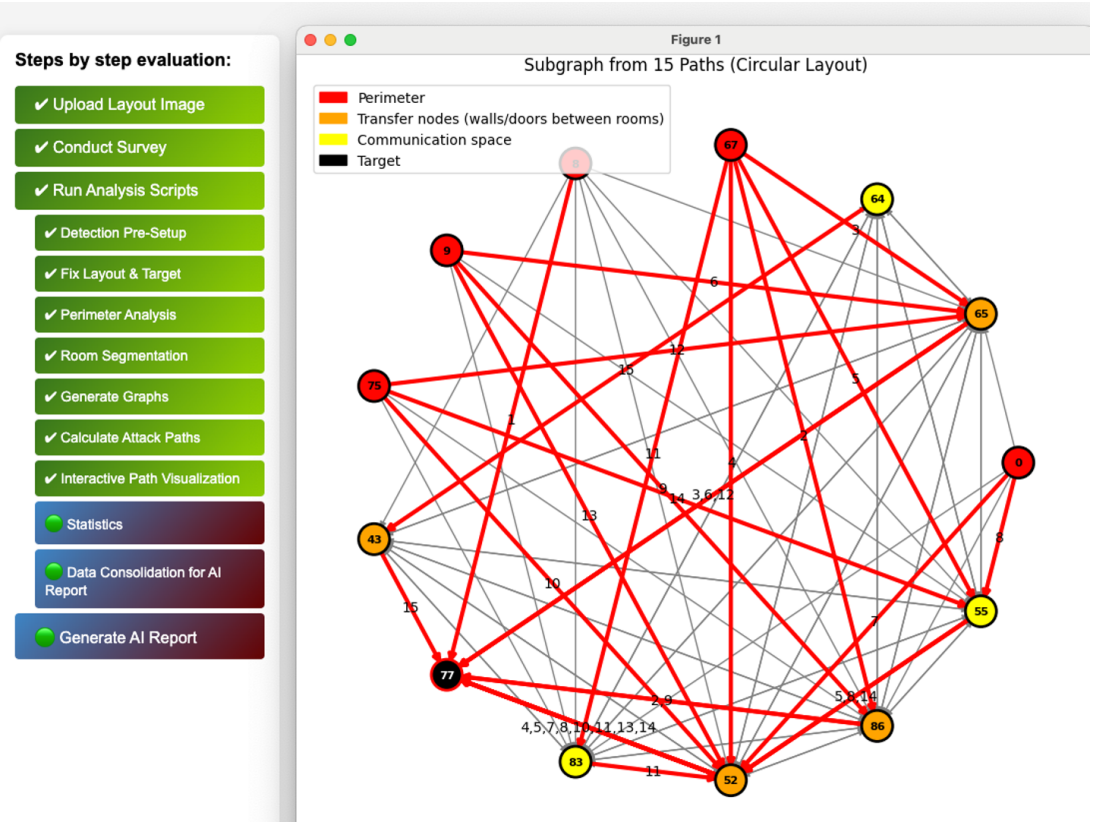
Interpretacja wyników statystycznych

Analiza statystyk pozwala na wyciągnięcie kluczowych wniosków:

Identyfikacja hot-spotów bezpieczeństwa

Węzły i krawędzie o najwyższej częstotliwości wykorzystania wskazują:

- Miejsca wymagające priorytetowego wzmocnienia zabezpieczeń
- Punkty, gdzie inwestycje w systemy detekcji przyniosą największy efekt
- Obszary krytyczne dla ogólnej odporności obiektu na ataki



Rysunek 6.10: Analiza statystyk ruchu - wizualizacja pierwszych 15 ścieżek ataków w układzie kołowym ze wskazaniem częstotliwości wykorzystania węzłów i krawędzi. Możemy zobaczyć, które fragmenty wydają się być najczęściej występujące w najbardziej efektywnych ścieżkach ataku.

Ocena redundancji systemu ochrony

- Węzły występujące w licznych ścieżkach wskazują na brak alternatywnych tras
- Wysoką koncentrację ruchu można interpretować jako słabość strukturalną
- Izolowane węzły mogą wskazywać na nadmiarowe zabezpieczenia

Optymalizacja alokacji zasobów

Statystyki umożliwiają:

- Racjonalne planowanie rozmieszczenia systemów bezpieczeństwa
- Priorytetyzację modernizacji w oparciu o rzeczywiste zagrożenia
- Optymalizację kosztów przez fokus na kluczowe punkty

Praktyczne zastosowania analizy

Planowanie systemów detekcji

Węzły o wysokiej częstotliwości wykorzystania powinny być wyposażone w:

- Redundantne systemy detekcji
- Czujki o podwyższonej czułości
- Dodatkowe systemy komunikacji alarmowej

Projektowanie barier fizycznych

Analiza wskazuje miejsca, gdzie:

- Wzmocnienie istniejących barier przyniesie największy efekt
- Dodanie nowych przeszkód może znacząco utrudnić atak
- Modyfikacja układu komunikacyjnego może poprawić bezpieczeństwo

Strategie operacyjne

Statystyki ruchu informują o:

- Optymalnym rozmieszczeniu personelu ochrony
- Częstotliwości patroli w kluczowych obszarach
- Priorytetach w czasie reagowania na alarmy

Analiza statystyk stanowi integralną część procesu oceny, przekształcając abstrakcyjne wyniki algorytmów w konkretne, praktyczne rekomendacje dla poprawy bezpieczeństwa obiektu. Możemy zobaczyć, które fragmenty wydają się być najczęściej występujące w najbardziej efektywnych ścieżkach ataku, co dostarcza kluczowych informacji dla strategicznego planowania zabezpieczeń.

6.4. Część IV: Synteza wyników z wykorzystaniem LLM

Końcowy etap procesu wykorzystuje duże modele językowe do integracji wszystkich wcześniejszych analiz w spójny, praktyczny raport. System konsoliduje dane z detekcji obrazu, analizy grafowej i Security Survey, przekształcając je w konkretne rekomendacje dotyczące wzmocnienia bezpieczeństwa obiektu.

Etap X: Konsolidacja danych i generowanie raportu AI

Ostatni etap procesu oceny polega na konsolidacji wszystkich zebranych danych i automatycznym generowaniu kompleksowego raportu przy wykorzystaniu modeli językowych (LLM). Na tym etapie system integruje wyniki wszystkich poprzednich analiz - od detekcji elementów architektonicznych, przez Security Survey, analizę ścieżek ataku, aż po statystyki częstotliwości wykorzystania węzłów - i przekształca je w strukturalny, czytelny raport dostarczający konkretnych rekomendacji.

Architektura systemu raportowania

System raportowania wykorzystuje zaawansowane modele językowe do analizy i syntezy danych z poprzednich etapów. Proces generowania raportu obejmuje:

Agregację danych wejściowych

- Wyniki analizy ścieżek ataku (174 przeanalizowane ścieżki)
- Statystyki wykorzystania węzłów i krawędzi
- Dane kontekstowe z Security Survey
- Parametry degradacji systemów bezpieczeństwa
- Charakterystyki wykrytych elementów architektonicznych

Analiza wzorców i trendów

LLM przeprowadza wielowymiarową analizę w celu:

- Identyfikacji kluczowych słabości w systemie ochrony
- Oceny względnej skuteczności poszczególnych barier
- Wykrycia korelacji między różnymi aspektami bezpieczeństwa
- Prognozowania trendów degradacji w czasie

Struktura generowanego raportu

Rysunek 6.11 przedstawia przykład automatycznie wygenerowanego raportu, który składa się z następujących sekcji:

Section 1: Attack Paths Analysis

Analiza ścieżek ataku zawiera kluczowe statystyki:

– Całkowita liczba przeanalizowanych ścieżek (174)

– Średni koszt ścieżki wymagający dotarcia do alternatywnego celu (739.36)

– Identyfikacja najczęściej wykorzystywanego typu barier (SingleDoor)

– Analiza pozycji bezpieczeństwa względem podatności

Section 2: Trend Analysis

Niestety, brak danych historycznych uniemożliwia analizę trendów w czasie rzeczywistym. System odnotowuje to ograniczenie i sugeruje gromadzenie danych w przyszłości.

Section 3: Detailed Physical Security Report

Szczegółowy raport bezpieczeństwa fizycznego, w tym:

– Ocena najbardziej narażonych typów barier

– Rekomendacje dotyczące wzmocnienia zabezpieczeń

– Analiza efektywności różnych systemów detekcji

Section 4: Elements Recommended for Reinforcement

Konkretne rekomendacje wzmocnienia:

– Priorytetyzacja elementów SingleDoor jako najczęściej wykorzystywanych

– Wzmocnienie zabezpieczeń drzwi i okien w kluczowych lokalizacjach

– Instalacja dodatkowych czujek w podatnych obszarach

Section 5: Security Enhancements

Proponowane usprawnienia bezpieczeństwa obejmujące zarówno rozwiązania techniczne, jak i organizacyjne.

Section 6: Summary and Urgent Actions

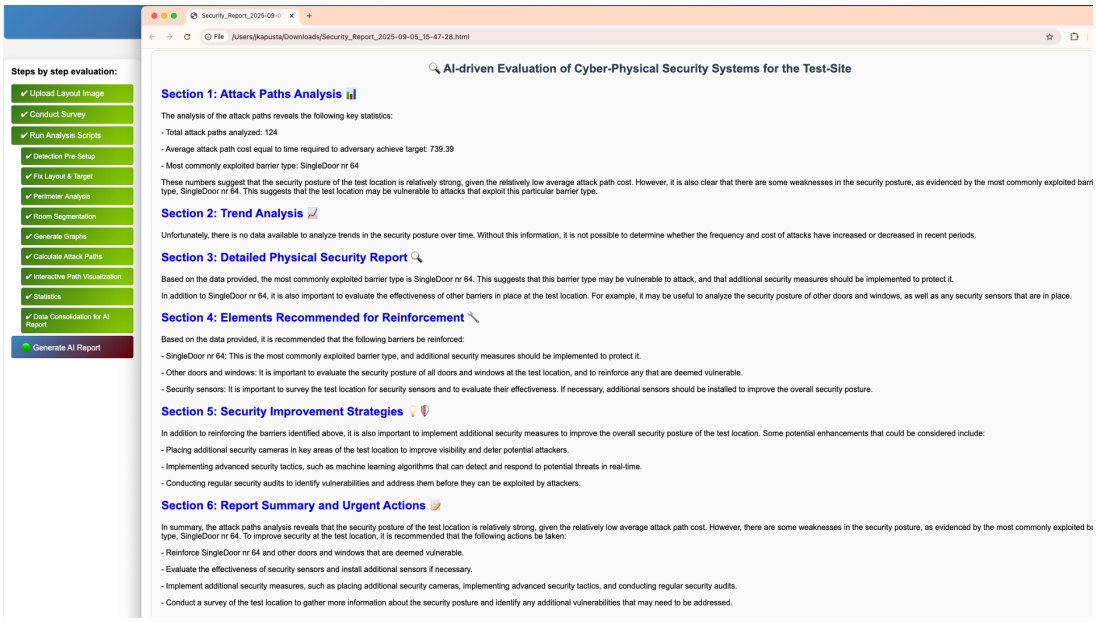
Podsumowanie z pilnymi działaniami, w tym:

– Wzmocnienie SingleDoor jako elementu o najwyższym priorytecie

– Ocena skuteczności czujek i instalacja dodatkowych systemów

– Implementacja zaawansowanych rozwiązań bezpieczeństwa

– Przeprowadzenie regularnych audytów



Rysunek 6.11: Automatycznie wygenerowany raport AI - kompleksowa analiza bezpieczeństwa obiektu z konkretnymi rekomendacjami opartymi na danych z wszystkich etapów oceny.

Wartość dodana automatycznego raportowania

Obiektywność analizy

Wykorzystanie LLM zapewnia:

- Brak subiektywnych preferencji w ocenie zagrożeń
- Systematyczne podejście do analizy wszystkich danych
- Spójność w metodologii oceny różnych aspektów bezpieczeństwa

Kompleksowość ujęcia

System uwzględnia:

- Interakcje między różnymi elementami systemu ochrony
- Wpływ degradacji na długoterminową skuteczność
- Kontekst operacyjny i specyfikę organizacji

Praktyczność rekomendacji

Generowane zalecenia charakteryzują się:

- Konkretnym odniesieniem do zidentyfikowanych problemów
- Priorytetyzacją działań według kryterium efektywności
- Uwzględnieniem ograniczeń budżetowych i operacyjnych

Integracja z modelem degradacji

Raport automatycznie uwzględnia model degradacji opisany w rozprawie:

- Prognozuje zmiany skuteczności systemów w czasie
- Rekomenduje harmonogramy konserwacji i modernizacji
- Wskazuje elementy wymagające priorytetowej wymiany
- Optymalizuje inwestycje z perspektywy długoterminowej

Automatyczne generowanie raportu stanowi kulminację całego procesu oceny, przekształcając złożone dane techniczne w praktyczne narzędzie wspomagające podejmowanie decyzji w zakresie bezpieczeństwa. System dostarcza nie tylko diagnozy aktualnego stanu, ale również konkretnych, opartych na danych rekomendacji działań mających na celu poprawę efektywności ochrony obiektu.

Podsumowanie rozdziału

Przedstawiony w niniejszym rozdziale proces implementacji narzędzia do oceny efektywności Cyber-Fizycznych Systemów Ochrony stanowi praktyczne zastosowanie założeń teoretycznych i metodologii opracowanej w rozprawie. System przechodzi przez dziesięć kluczowych etapów - od akwizycji planów budynku, przez Security Survey i detekcję elementów architektonicznych, aż po automatyczne generowanie kompleksowych raportów z wykorzystaniem modeli językowych. Narzędzie, choć wymaga dalszych usprawnień, już na tym etapie jest pełnoprawnym narzędziem w rękach specjalisty, który rozumie jego zasady działania. Kluczowe mocne strony implementacji obejmują:

- **Systematyczne podejście**: uporządkowany proces analizy zapewnia spójność i powtarzalność wyników
- **Integrację różnych źródeł danych**: od automatycznej detekcji po dane kontekstowe z Security Survey
- **Transparentność procesu**: każdy etap dostarcza czytelnych wyników umożliwiających weryfikację
- **Praktyczne zastosowanie teorii**: implementacja modelu degradacji w rzeczywistych scenariuszach oceny
- **Interaktywność**: możliwość manualnej korekty i walidacji wyników na kluczowych etapach

Szczególnie istotnym aspektem jest filozofia projektowa, zgodnie z którą **narzędzie nie zastępuje eksperta, a stanowi istotne wsparcie w jego zadaniach**. Ta zasada przejawia się w kilku kluczowych elementach:

Kontrola eksperta nad procesem

- Możliwość manualnej korekty wyników automatycznej detekcji
- Interaktywna wizualizacja umożliwiająca walidację logiki systemu
- Przejrzysty dostęp do danych wejściowych i założeń każdego etapu

Wspomaganie, nie zastępowanie decyzji

- System dostarcza analiz i rekomendacji, ale ostateczne decyzje pozostają w gestii eksperta
- Wizualizacje i statystyki służą jako narzędzia wspomagające zrozumienie problemu
- Automatyczne raporty stanowią punkt wyjścia do dalszej analizy eksperckiej

Redukcja nakładu pracy przy zachowaniu jakości

- Automatyzacja rutynowych zadań (detekcja, generowanie grafów, statystyki)
- Znaczące skrócenie czasu analizy w porównaniu z metodami manualnymi
- Zwiększenie dokładności przez eliminację błędów ludzkich w obliczeniach

Implementacja potwierdza praktyczną wykonalność koncepcji przedstawionych w części teoretycznej rozprawy. Model degradacji, integracja aspektów cyber-fizycznych oraz metodologia oceny względnie racjonalnych ścieżek ataku znajdują odzwierciedlenie w funkcjonalnym narzędziu, które może być wykorzystywane w rzeczywistych scenariuszach oceny bezpieczeństwa. Obszary wymagające dalszego rozwoju obejmują przede wszystkim rozszerzenie bazy danych treningowych dla algorytmów YOLO, implementację bardziej zaawansowanych modeli degradacji opartych na danych rzeczywistych oraz integrację z zewnętrznymi systemami zarządzania bezpieczeństwem. Niemniej jednak, już w obecnej formie narzędzie stanowi znaczący krok naprzód w automatyzacji i systematyzacji procesów oceny efektywności systemów ochrony fizycznej. Zaprezentowany system implementacyjny dowodzi, że zaawansowane metodologie teoretyczne mogą być skutecznie przekształcone w praktyczne narzędzia wspomagające ekspertów bezpieczeństwa, pod warunkiem zachowania odpowiedniej równowagi między automatyzacją a kontrolą ludzką nad procesem decyzyjnym.

7. Konkluzje i analiza krytyczna

7.1. Główne osiągnięcia badawcze

Rozprawa ta zaproponowała zintegrowane i dynamiczne podejście do oceny efektywności Cyber-Fizycznych Systemów Ochrony (CPPS), łącząc perspektywę probabilistyczną, uczenie maszynowe oraz wykorzystanie dużych modeli językowych (LLM). Opracowany zestaw metod i narzędzi badawczych wskazuje nowe kierunki rozwoju zarówno w teorii, jak i w praktyce inżynierii bezpieczeństwa.

W toku badań zrealizowaliśmy kilka istotnych zadań, które stanowią nasz wkład w dziedzinę. Wydaje się, że po raz pierwszy w literaturze przedmiotu zaproponowaliśmy probabilistyczny model degradacji specyficzny dla Elektronicznych Systemów Bezpieczeństwa (ESS), uwzględniający pięć kluczowych czynników: starzenie się komponentów, wzrost wiedzy przeciwnika, rozwój technologiczny narzędzi atakujących, pojawianie się podatności oraz działania konserwacyjne. Symulacje Monte Carlo wykonane dla $N = 10{,}000$ przebiegów sugerują znaczący spadek efektywności systemów w czasie, z możliwym spadkiem poniżej 50% skuteczności w horyzoncie 6,5-7 lat.

Klasyczny model EASI został uzupełniony o wskaźniki degradacji, co pozwala na bardziej realistyczne odwzorowanie efektywności CPPS w długim horyzoncie czasowym. Przeprowadzone analizy wskazują na potencjalne obniżenie estymowanych wskaźników skuteczności o 15-25 punktów procentowych po pięciu latach eksploatacji w porównaniu z klasycznym podejściem. Ta różnica ma fundamentalne znaczenie dla planowania budżetowego i decyzji o modernizacji systemów ochronnych.

Metodologii Contextual Object Grouping (COG) umożliwia automatyczną interpretację diagramów bezpieczeństwa, osiągając wyniki na poziomie 91,7% mAP@0.50-0.95 w detekcji elementów architektonicznych. COG stanowi skuteczny pomost między analogowymi planami a cyfrowymi reprezentacjami systemów bezpieczeństwa, umożliwiając cyfrową rekonstrukcję struktury ochronnej z wysoką precyzją. Podejście to rozwiązuje fundamentalny problem braku standaryzacji symboliki na planach bezpieczeństwa poprzez dynamiczne uczenie się kontekstu z legend diagramów.

Zaproponowaliśmy i przetestowaliśmy heurystyczną analizę ścieżek ataku wykorzystującą algorytm A* z uwzględnieniem kontekstu operacyjnego i degradacji. Metoda ta oferuje wspomaganą komputerowo analizę kompleksowych scenariuszy zagrożeń, redukując teoretycznie nieskończoną liczbę możliwych ścieżek do zbioru praktycznie analizowanego. Zastosowanie funkcji heurystycznej $f(P_d, P_{ce}, T_{dl}, T_{rr})$ umożliwia priorytetyzację najbardziej prawdopodobnych i niebezpiecznych tras intruzji.

Duże modele językowe wykorzystaliśmy do konsolidacji, raportowania i interpretacji wyników analiz, co znacząco wspiera praktyczne zastosowania modelu w środowisku eksperckim. Automatyczne ge-

nerowanie spersonalizowanych raportów i rekomendacji konserwacyjnych przełamuje barierę między zaawansowanymi analizami technicznymi a praktycznymi potrzebami operatorów systemów bezpieczeństwa.

Integracja metodologiczna

Szczególną wartością przedstawionej pracy jest integracja różnych metodologii w spójne podejście analityczne. Model degradacji oparty na symulacjach Monte Carlo łączy się z klasyczną metodologią EASI, tworząc rozszerzoną wersję EASI-Δ, która uwzględnia czasową zmienność parametrów systemu. Metodologia COG automatyzuje proces ekstrakcji informacji z planów architektonicznych, dostarczając danych dla algorytmów przeszukiwania grafów. Duże modele językowe działają jako warstwa interpretacyjna, przekładając wyniki techniczne na rekomendacje zrozumiałe dla decydentów.

Z punktu widzenia praktyki w obszarze bezpieczeństwa fizycznego, integracja ta umożliwia kompleksowy proces oceny CPPS, który rozpoczyna się od automatycznej analizy planów architektonicznych, przechodzi przez modelowanie degradacji i symulację scenariuszy ataku, a kończy się na generowaniu raportów i rekomendacji konserwacyjnych. Całość procesu może być przeprowadzona w znacznej mierze automatycznie, przy zachowaniu możliwości weryfikacji i korekty przez ekspertów.

7.2. Analiza krytyczna ograniczeń

Parametryzacja i świadome uproszczenia

Największym ograniczeniem przedstawionego modelu jest jego parametryzacja. Wartości parametrów zostały dobrane w oparciu o dostępną literaturę, konsultacje eksperckie i badania symulacyjne, a nie na podstawie obszernych danych empirycznych z rzeczywistych instalacji CPPS. Rozkłady probabilistyczne mogą nie w pełni odzwierciedlać rzeczywistość operacyjną różnych typów instalacji, a częstotliwość występowania krytycznych podatności oparto na danych historycznych z okresu 2015-2020, które mogą okazać się nieadekwatne do przyszłych trendów w cyberbezpieczeństwie.

Efekty konserwacji zostały ujęte w modelu w sposób równomierny i systematyczny, co nie zawsze odpowiada zróżnicowanym praktykom organizacyjnym. Niektóre organizacje przeprowadzają konserwację częściej lub rzadziej niż co kwartał, a jakość i zakres działań konserwacyjnych mogą się znacznie różnić. Model nie uwzględnia również korelacji między czynnikami degradacji, co może prowadzić do niedoszacowania lub przeszacowania rzeczywistych efektów synergistycznych.

Przyjęte uproszczenia, takie jak addytywność czynników degradacji, ich statystyczna niezależność czy pominięcie czynników środowiskowych (wilgotności, temperatury, intensywności użytkowania), pozwoliły na utrzymanie przejrzystości i implementowalności modelu, lecz prawdopodobnie kosztem jego pełnego realizmu. Te świadome kompromisy były konieczne dla umożliwienia formalizacji i implementacji kompleksowego podejścia na obecnym etapie rozwoju dziedziny.

Ograniczenia technologiczne Metodologia COG

Metodologia COG, mimo osiągnięcia wysokich wskaźników skuteczności detekcji, wykazuje pewne ograniczenia techniczne. Model został trenowany na zbiorze danych ograniczonym do siedmiu podstawowych klas architektonicznych, co może być niewystarczające dla kompleksowych analiz bezpieczeństwa w różnorodnych środowiskach. Różnice w standardach rysunkowych, jakości skanów i skali diagramów mogą wpływać na skuteczność detekcji.

COG wykazuje kontekstową świadomość, wykrywając symbole jedynie w kontekście legend, ale nie w głównych obszarach diagramów. To zachowanie, choć zgodne z założeniami metodologicznymi, może ograniczać zastosowanie do analiz wymagających identyfikacji wszystkich wystąpień symboli bez względu na kontekst. Dodatkowo, model wymaga post-processingu dla eliminacji duplikatów i walidacji geometrycznej, co zwiększa złożoność implementacyjną.

Ograniczenia heurystycznej analizy ścieżek

Algorytm przeszukiwania ścieżek A* z zaproponowaną funkcją heurystyczną $f(P_d, P_{ce}, T_{dl}, T_{rr})$ może nie uwzględniać wszystkich aspektów rzeczywistych scenariuszy ataku. Funkcja heurystyczna została zdefiniowana na podstawie parametrów klasycznej metodologii EASI, co może nie obejmować nowoczesnych technik ataku cybernetycznego lub socjotechnicznego. Model zakłada również racjonalność przeciwnika i dążenie do optymalizacji ścieżki, co nie zawsze odpowiada rzeczywistości.

Przeszukiwanie ogranicza się do reprezentacji dwuwymiarowej budynków, co jest niewystarczające dla kompleksowej analizy obiektów wielokondygnacyjnych. Brak uwzględnienia czasowej zmienności obsady obiektów, systemów kontroli dostępu w czasie rzeczywistym czy dynamicznych procedur bezpieczeństwa może prowadzić do błędnych ocen realnych zagrożeń.

Świadomość ograniczeń jako metodologiczny fundament

Wszystkie zidentyfikowane ograniczenia należy traktować nie jako fundamentalne słabości, lecz jako świadome punkty wyjścia dla dalszych badań empirycznych. Model został zaprojektowany jako metodologiczne ramy (framework), które mogą być sukcesywnie rozwijane i dostrajane w miarę gromadzenia danych operacyjnych. Przyjęte uproszczenia są nieuniknione na obecnym etapie rozwoju dziedziny i umożliwiają stworzenie podstaw do bardziej zaawansowanych implementacji.

Zidentyfikowane ograniczenia wyznaczają konkretne kierunki przyszłych prac, które powinny koncentrować się na systematycznym pozyskiwaniu danych z rzeczywistych instalacji CPPS, empirycznej kalibracji parametrów dla różnych typów systemów, modelowaniu zależności nieliniowych między czynnikami degradacji oraz uwzględnianiu środowiskowych, organizacyjnych i ekonomicznych determinant skuteczności systemów.

7.3. Wkład w stan wiedzy

Nowatorski charakter podejścia

Praca wnosi do literatury przedmiotu kilka istotnych elementów metodologicznych i technicznych. Wydaje się, że jako pierwsi zaproponowaliśmy probabilistyczny model degradacji specyficzny dla CPPS, uwzględniający jednocześnie czynniki techniczne (starzenie komponentów), środowiskowe (ewolucja zagrożeń) i operacyjne (działania konserwacyjne). Praktyczne rozszerzenie uznanej metodologii EASI o wskaźniki degradacji i cyberodporności stanowi znaczący postęp w kierunku realistycznej oceny długoterminowej efektywności systemów ochronnych.

Integracja Metodologia COG z heurystyczną analizą ścieżek ataku tworzy kompletny proces od automatycznej analizy diagramów do generowania scenariuszy zagrożeń. Zastosowanie LLM do automatycznej interpretacji i raportowania w kontekście analizy bezpieczeństwa otwiera nowe możliwości komunikacji między systemami technicznymi a użytkownikami końcowymi.

Metodologia COG wprowadza koncepcję klasy kontekstowej COG jako pośredniego poziomu ontologicznego między atomową detekcją obiektów a interpretacją semantyczną. To podejście umożliwia modelom bezpośrednie postrzeganie kontekstu zamiast wnioskowania o nim poprzez post-processing, co stanowi fundamentalną zmianę paradygmatu w analizie diagramów technicznych.

Interdyscyplinarny charakter rozwiązania

Zaproponowany model i narzędzia badawcze łączą osiągnięcia z kilku dziedzin, tworząc spójną metodologię interdyscyplinarną. Z inżynierii niezawodności zaczerpnięto modelowanie procesów degradacji i analizy cyklu życia, z cyberbezpieczeństwa - analizę ewolucji zagrożeń i podatności, ze sztucznej inteligencji - zastosowanie uczenia maszynowego i LLM, z teorii grafów - modelowanie ścieżek ataku i optymalizację, a z zarządzania bezpieczeństwem - integrację z praktykami operacyjnymi.

Takie interdyscyplinarne podejście umożliwia zastosowanie proponowanych metod w różnych sektorach infrastruktury krytycznej, od energetyki, przez transport, po sektor finansowy. Uniwersalność metodologii wynika z abstrakcyjnego potraktowania CPPS jako systemów o określonych właściwościach probabilistycznych, niezależnie od konkretnej domeny zastosowań.

Potencjalne implikacje praktyczne

Rezultaty badań wskazują na możliwą potrzebę rewizji standardowych, dziesięcioletnich cykli wymiany systemów ochronnych w kierunku bardziej elastycznych, opartych na rzeczywistym stanie degradacji. Wdrożenie dynamicznych harmonogramów konserwacji dostosowanych do przewidywanego tempa degradacji może znacząco poprawić efektywność kosztową utrzymania systemów bezpieczeństwa.

Szeroka integracja oceny CPPS z narzędziami AI dla wsparcia predykcyjnego utrzymania może prowadzić do powstania nowej generacji inteligentnych systemów bezpieczeństwa zdolnych do samodiagnostyki i adaptacji. Rozwój cyfrowych bliźniaków jako platform walidacji i testowania różnych strategii

utrzymania może istotnie wpłynąć na podejście do planowania bezpieczeństwa w infrastrukturze krytycznej.

Implementacja systemów ciągłego monitorowania efektywności zamiast okresowych przeglądów może doprowadzić do fundamentalnej zmiany w kulturze bezpieczeństwa organizacji, przesuwając akcent z reaktywnego na proaktywne zarządzanie ryzykiem.

7.4. Kierunki dalszych badań

Walidacja empiryczna jako fundament

Najważniejszym i najbardziej pilnym wyzwaniem jest systematyczna empiryczna walidacja modelu na danych z rzeczywistych instalacji CPPS. Ten kompleksowy proces będzie wymagał nawiązania partnerstw strategicznych z operatorami infrastruktury krytycznej w różnych sektorach (energetyka, transport, finanse, ochrona zdrowia), budowania relacji z instytucjami badawczymi i uniwersytetami mającymi dostęp do danych operacyjnych oraz partnerstwa z producentami systemów bezpieczeństwa i integratorami.

Rozwój infrastruktury badawczej powinien obejmować opracowanie standardów i protokołów gromadzenia danych operacyjnych zachowujących wymagania bezpieczeństwa, stworzenie bezpiecznych mechanizmów udostępniania danych wrażliwych dla celów badawczych poprzez szyfrowanie, anonimizację i kontrolę dostępu, oraz rozwój narzędzi do automatycznego gromadzenia i przetwarzania danych z różnych źródeł, takich jak sensory, logi systemowe i raporty serwisowe.

Długoterminowe programy monitorowania powinny obejmować organizację wieloletnich programów obserwacji degradacji w różnych warunkach środowiskowych, systematyczne dokumentowanie efektów różnych strategii konserwacyjnych i ich wpływu na tempo degradacji, tworzenie kohort systemów o podobnych charakterystykach dla analiz porównawczych oraz integrację z istniejącymi programami monitorowania infrastruktury.

Zaawansowane rozszerzenia technologiczne

Kolejny etap rozwoju powinien obejmować integrację z nowoczesnymi technologiami i metodami obliczeniowymi. Zastosowanie głębokich sieci neuronowych do wykrywania złożonych wzorców degradacji w danych wielowymiarowych, wykorzystanie uczenia przez wzmacnianie do optymalizacji strategii konserwacyjnych w dynamicznych środowiskach, oraz implementacja modeli attention-based do identyfikacji kluczowych czynników wpływających na degradację mogą znacząco poprawić dokładność predykcji.

Internet rzeczy (IoT) i analityka brzegowa otwierają możliwości integracji z czujnikami IoT nowej generacji, umożliwiającymi ciągłe monitorowanie parametrów środowiskowych i operacyjnych, implementacji analityki brzegowej dla real-time'owej detekcji anomalii i predykcji degradacji, rozwoju inteligentnych sensorów zdolnych do samodzielnej analizy i raportowania oraz integracji z systemami zarządzania budynkami (BMS) i infrastrukturą (IMS).

Technologie rozproszone i blockchain mogą zostać wykorzystane do eksperymentów z technologią blockchain dla zapewnienia niezmienności i audytowalności danych pomiarowych, rozwoju systemów

rozproszonych umożliwiających bezpieczne udostępnianie danych między organizacjami, implementacji smart contracts dla automatyzacji procesów konserwacyjnych oraz zastosowania technologii zero-knowledge proof dla prywatnej walidacji modeli.

Rozwój regulacyjny i standaryzacyjny

Praktyczne wdrożenie będzie wymagać znaczących zmian w środowisku regulacyjnym i standaryzacyjnym. Opracowanie elastycznych ram prawnych uwzględniających wskaźniki degradacji i ryzyka w wymaganiach regulacyjnych, rozwój procedur certyfikacji systemów CPPS opartych na ciągłej ocenie efektywności, integracja wyników modeli predykcyjnych z procesami licencyjnymi i audytowymi oraz harmonizacja wymagań między różnymi jurysdykcjami i sektorami będą konieczne dla szerokiego zastosowania proponowanych metod.

Rozwój standardów ISO/IEC dla modelowania degradacji CPPS, harmonizacja metodologii między organizacjami takimi jak NIST, ENISA, i narodowymi centrami cyberbezpieczeństwa, opracowanie protokołów wymiany danych i interoperacyjności między różnymi systemami analitycznymi, oraz tworzenie certyfikowanych procedur walidacji i kalibracji modeli stanowią priorytetowe obszary standaryzacji.

Aspekty etyczne i prywatności wymagają rozwoju ram etycznych dla wykorzystania AI w systemach bezpieczeństwa, standardów ochrony prywatności dla danych operacyjnych infrastruktury krytycznej, protokołów transparentności i odpowiedzialności dla systemów AI w zastosowaniach krytycznych oraz mechanizmów kontroli społecznej i nadzoru demokratycznego nad systemami automatycznymi.

Transformacja organizacyjna i kompetencyjna

Skuteczne wdrożenie będzie wymagać fundamentalnych zmian w organizacjach i rozwoju nowych kompetencji. Rozwój kultury proaktywnej konserwacji opartej na danych w organizacjach zarządzających infrastrukturą krytyczną, implementacja systemów motywacyjnych zachęcających do adopcji nowych technologii i metod, tworzenie interdyscyplinarnych zespołów łączących kompetencje techniczne, analityczne i zarządcze, oraz rozwój procedur zarządzania zmianą dla stopniowego wdrażania nowych podejść będą kluczowe dla sukcesu transformacji.

Opracowanie specjalistycznych programów kształcenia dla inżynierów bezpieczeństwa w zakresie modelowania predykcyjnego, szkolenia dla personelu operacyjnego w interpretacji wyników modeli AI i podejmowaniu decyzji opartych na danych, programy dla kadry zarządczej dotyczące strategicznego wykorzystania analiz predykcyjnych, oraz rozwój certyfikacji profesjonalnych dla specjalistów od inteligentnych systemów bezpieczeństwa wymagają systematycznego podejścia do budowania kompetencji.

7.5. Podsumowanie

Charakter wkładu

Prezentowany model i metody należy traktować jako metodologiczny fundament dla dalszych badań nad inteligentnymi systemami bezpieczeństwa. Mając pełną świadomość ograniczeń i stosowanych

uproszczeń, można je uznać za świadome ramy początkowe, które nie tylko wyznaczają obiecujący kierunek badań, lecz także precyzyjnie wskazują obszary, gdzie konieczne jest systematyczne gromadzenie danych empirycznych.

Wartość dodana pracy polega nie tyle na dostarczeniu gotowego rozwiązania, ile na sformalizowaniu problemu degradacji CPPS w sposób umożliwiający dalsze badania empiryczne, zaproponowaniu konkretnych narzędzi i metod, które mogą być rozwijane i dostrajane, zidentyfikowaniu kluczowych czynników wymagających monitorowania i analizy, oraz wskazaniu interdyscyplinarnych możliwości integracji różnych dziedzin wiedzy.

Uwarunkowania praktycznego zastosowania

Praktyczna wartość proponowanego podejścia będzie zależała od zdolności adaptacji - możliwości dostosowania modeli do specyfiki różnych typów instalacji i warunków operacyjnych, jakości i dostępności danych - rozwoju systemów gromadzenia i udostępniania danych operacyjnych przy zachowaniu wymogów bezpieczeństwa, gotowości organizacyjnej - akceptacji przez instytucje zaawansowanych strategii dynamicznego zarządzania utrzymaniem, postępu technologicznego - dalszego rozwoju w dziedzinie AI, IoT i metod probabilistycznych, oraz wsparcia regulacyjnego - włączenia wyników modeli predykcyjnych do formalnych ram prawnych i standardów branżowych.

Wizja perspektywiczna

Rozprawa może stanowić istotny krok ku rozwojowi inteligentnych systemów CPPS, które będą charakteryzować się zdolnością do samodiagnostyki - automatycznego wykrywania oznak degradacji i przewidywania awarii, predykcji adaptacyjnej - dostosowania modeli predykcyjnych do zmieniających się warunków, optymalizacji w czasie rzeczywistym - ciągłego dostrajania strategii utrzymania, oraz współpracy między systemami - koordynacji działań w ramach całej infrastruktury krytycznej.

Realizacja tej wizji będzie wymagać nie tylko dalszego rozwoju technologicznego, ale także fundamentalnych zmian w podejściu do zarządzania bezpieczeństwem - przejścia od reaktywnych modeli utrzymania do proaktywnych, inteligentnych systemów zdolnych do antycypowania i zapobiegania degradacji przed jej krytycznymi konsekwencjami.

Przedstawiona praca może być postrzegana jako próba zdefiniowania nowego paradygmatu w inżynierii bezpieczeństwa - paradygmatu, w którym systemy ochrony nie są traktowane jako statyczne infrastruktury, ale jako dynamiczne, uczące się organizmy zdolne do ciągłej adaptacji i samodoskonalenia. Ten paradygmat, wspierany przez postęp w dziedzinie sztucznej inteligencji i analityki predykcyjnej, może fundamentalnie zmienić sposób projektowania, wdrażania i utrzymania systemów bezpieczeństwa infrastruktury krytycznej.

A. Dodatek A - Raport LLM

Security Analysis Report for Test_Facility, Kraków, Poland

Author:.....

Date:.....

Section 1: Attack Paths Analysis

- **Total Attack Paths Analyzed:** 124
- **Average Attack Path Cost (Time):** 12.32 minutes (739.39 seconds)
- **Most Commonly Exploited Barrier Type:** Internal Walls (IntWall)

Analysis:
The high number of attack paths indicates multiple potential entry points and vulnerabilities within the facility. The average time required for an adversary to breach the facility is **12.32 minutes**, which is dangerously close to the **15-minute response time** of the security team. This suggests that adversaries could achieve their objectives before security arrives, particularly if they exploit internal walls. The frequent targeting of internal walls highlights a critical weakness in the facility's internal security structure.

Section 2: Trend Analysis

- **Observed Trends:**
  Attack paths often involve internal walls, suggesting that adversaries prioritize exploiting weaknesses in internal security once they bypass perimeter defenses. The average breach time of **12.32 minutes** is concerning, as it leaves only a **2.68-minute window** for security to respond before an adversary succeeds.
- **Changes Over Time:**
  If historical data were available, it would be essential to analyze whether breach attempts have become faster or more frequent. The current data suggests a consistent pattern of targeting internal barriers, which could worsen if left unaddressed.

Section 3: Detailed Physical Security Report

- **Weaknesses in Most Exploited Barriers:**
  Internal Walls (IntWall): These are frequently exploited, indicating potential weaknesses in access control, structural integrity, or monitoring. Adversaries may be leveraging gaps in surveillance or physical vulnerabilities in these walls.

- **Vulnerabilities:**
  Internal walls may lack sufficient reinforcement or real-time monitoring, making them easier to breach once an adversary gains access. Windows and single doors are also targeted, suggesting these points require additional security layers.

- **Current Security Countermeasures:**
  The facility is equipped with **access control, alarm systems, and CCTV**, monitored 24/7 by a commercial security firm. The **15-minute response time** for armed intervention is a strength, but the **12.32-minute average breach time** means adversaries could succeed before security arrives if they exploit internal weaknesses.

Section 4: Elements Recommended for Reinforcement

- **Barriers Requiring Reinforcement:**
  Internal Walls (IntWall): Strengthen with reinforced materials, additional access control, or intrusion detection systems.
  Windows and Single Doors: Install shatterproof glass, reinforced door frames, and alarms to delay breaches.

- **Security Zones Needing Attention:**
  High-Traffic Internal Areas: Increase sensor coverage and surveillance in zones where internal walls are frequently breached.
  Perimeter Adjacent to Internal Walls: Enhance monitoring to detect early signs of intrusion and slow down adversaries.

Section 5: Security Improvement Strategies

- **Enhancements for Sensor Placement:**
  Deploy additional motion sensors and surveillance cameras near internal walls and high-risk zones. Integrate **AI-driven anomaly detection** to identify suspicious activity patterns in real time.

- **Cost-Effective Reinforcements:**
  Access Control Upgrades: Implement multi-factor authentication for internal doors to slow down adversaries.
  Physical Barriers: Use bollards, reinforced doors, or barriers to restrict access to vulnerable internal areas.

- **Survey Results:**
  The survey indicates that **70% of security sensors are functioning optimally**, but **30% require maintenance or replacement**. Addressing this gap will improve detection capabilities and potentially increase the time required for a breach.

Section 6: Report Summary and Urgent Actions

- **Key Findings:**
  Internal walls are the primary target for adversaries, and the **12.32-minute average breach time** is too close to the **15-minute security response time**. While perimeter security is robust, internal defenses require immediate reinforcement to prevent adversaries from succeeding before security arrives.

- **Final Security Assessment:**
  The facility's perimeter security is strong, but internal vulnerabilities pose a significant risk. Reinforcing internal walls, windows, and doors, along with addressing sensor maintenance, will extend the time required for a breach and improve overall security.

- **Urgent Actions:**

  1. Reinforce internal walls with stronger materials, access control, and real-time monitoring.
  2. Upgrade or replace non-functional sensors to ensure full coverage of high-risk areas.
  3. Conduct regular security drills to test response times and identify further weaknesses.
  4. Reduce the average breach time below the 15-minute response threshold by implementing delays (e.g., reinforced doors, alarms) in critical paths.

Local Crime Insight

- **Crime Rate:** Moderate

- **Common Crimes:** Theft, vandalism, and unauthorized access are prevalent in the vicinity.

- **Security Incidents:** Occasional reports of attempted break-ins emphasize the need for proactive security measures.

Geopolitical Related Aspects

- **Stability:** Kraków is geopolitically stable, with a low risk of targeted attacks due to regional tensions.

- **Regional Security:** Collaborative measures with local law enforcement can further enhance security and response coordination.

Bibliografia

[1] Rada Unii Europejskiej, "Dyrektywa 2008/114/we z dnia 8 grudnia 2008 r. w sprawie identyfikacji i wyznaczenia europejskiej infrastruktury krytycznej oraz oceny konieczności poprawy jej ochrony," 2008. Dz.Urz. UE L 345 z 23.12.2008, str. 75–82.

[2] A. Zhang, "Cyber security in critical infrastructure," *Journal of Cyber Studies*, vol. 1, pp. 1–10, 2021.

[3] J. Kapusta, W. Bauer, and J. Baranowski, "Evaluation of the effectiveness of physical protection systems with consideration of its cyber-resilience," in *2023 27th International Conference on Methods and Models in Automation and Robotics (MMAR)*, pp. 457–461, 2023.

[4] European Union Agency for Cybersecurity (ENISA), "Threat landscape for critical sectors and services," 2021.

[5] J. Kapusta, W. Bauer, and J. Baranowski, "Ewaluacja skuteczności cyber-fizycznych systemów bezpieczeństwa (cpps) z uwzględnieniem degradacji elektronicznych systemów zabezpieczeń (ess)," *Automatyka i Robotyka AGH*, 2025. W druku.

[6] J. Kapusta, W. Bauer, and J. Baranowski, "A stochastic approach to modeling long-term degradation of the electronic security systems as a crucial element of evaluation of physical protection systems (pps)," in *IECON 2024 - 50th Annual Conference of the IEEE Industrial Electronics Society*, pp. 1–7, 2024.

[7] M. L. Garcia, *Design and Evaluation of Physical Protection Systems*. Elsevier, 2008.

[8] H. Bennett, "Easi approach to physical security evaluation," Tech. Rep. SAND–76-0500, United States, 1977.

[9] J. Clem, W. Atkins, and V. Urias, "Investigation of cyber-enabled physical attack scenarios," tech. rep., Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2015.

[10] S. Porter, "Physical protection systems and the cyber security component," Tech. Rep. LLNL-CONF-674448, Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), 2015.

[11] J. Depoy, J. Phelan, P. Sholander, B. Smith, G. Varnado, and G. Wyss, "Risk assessment for physical and cyber attacks on critical infrastructures," in *MILCOM 2005-2005 IEEE Military Communications Conference*, IEEE, 2005.

[12] M. Snell and J. Rivers, "The treatment of blended attacks in nuclear security effectiveness assessments," tech. rep., Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2015.

[13] Y. Guo, A. Yan, and J. Wang, "Cyber security risk analysis of physical protection systems of nuclear power plants and research on the cyber security test platform using digital twin technology," in *2021 International Conference on Power System Technology (POWERCON)*, IEEE, 2021.

[14] Z. Bowen, Y. Ming, Y. Hidekazu, and L. Hongxing, "Evaluation of physical protection systems using an integrated platform for analysis and design," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 11, pp. 2945–2955, 2017.

[15] D. Andiwijayakusuma, T. Setiadipura, A. Purqon, and Z. Su'ud, "The development of easi-based multi-path analysis code for nuclear security system with variability extension," *Nuclear Engineering and Technology*, vol. 54, no. 10, pp. 3604–3613, 2022.

[16] S. Narale, A. Verma, and S. Anand, "Structure and degradation of aluminum electrolytic capacitors," in *2019 National Power Electronics Conference, NPEC 2019*, pp. 1–6, 2019.

[17] C. Yu, L. Jiang, and J. Yuan, "Study of performance degradations in dc–dc converter due to hot carrier stress by simulation," *Microelectronics Reliability*, vol. 46, no. 9-11, pp. 1840–1843, 2006.

[18] M. Alam and F. Khan, "Reliability analysis and performance degradation of a boost converter," in *2013 IEEE Energy Conversion Congress and Exposition, ECCE 2013*, pp. 5592–5597, 2013.

[19] H. Huang, A. Boyer, and S. Dhia, "Analysis and modeling of passive device degradation for a long-term electromagnetic emission study of a dc-dc converter," *Microelectronics Reliability*, vol. 55, no. 9-10, pp. 2061–2066, 2015.

[20] D.-B. Du, J.-X. Zhang, Z.-J. Zhou, X.-S. Si, and C.-H. Hu, "Estimating remaining useful life for degrading systems with large fluctuations," *Journal of Control Science and Engineering*, 2018.

[21] G. Fang, R. Pan, and Y. Hong, "Copula-based reliability analysis of degrading systems with dependent failures," *Reliability Engineering and System Safety*, vol. 193, 2020.

[22] H. Dong, Z. Ding, and S. Zhang, *Deep reinforcement learning: Fundamentals, research and applications*. Springer, 2020.

[23] C. Hua, *Reinforcement Learning Aided Performance Optimization of Feedback Control Systems*. Springer, 2021.

[24] Y. Peng, S. Zhao, and H. Wang, "A digital twin based estimation method for health indicators of dc–dc converters," *IEEE Transactions on Power Electronics*, vol. 36, no. 2, pp. 2105–2118, 2021.

[25] S. Chen, S. Wang, P. Wen, and S. Zhao, "Digital twin for degradation parameters identification of dc–dc converters based on bayesian optimization," in *2021 IEEE International Conference on Prognostics and Health Management, ICPHM 2021*, pp. 1–9, 2021.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[27] J. Kapusta, W. Bauer, and J. Baranowski, "Contextual object grouping (cog): A novel framework for dynamic symbol interpretation in technical security diagrams," *Sensors*, vol. 25, no. 1, 2025. Submitted.

[28] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[32] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[33] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," *arXiv preprint arXiv:2401.17270*, 2024.

[34] B. Xiao, H. Wu, and Y. Wei, "Florence-2: Advancing a unified representation for a variety of vision tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4478–4488, 2024.

[35] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "Layoutlm: Pre-training of text and layout for document image understanding," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1192–1200, 2020.

[36] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "Layoutlmv3: Pre-training for document ai with unified text and image masking," *arXiv preprint arXiv:2204.08387*, 2022.

[37] H. Li, W. Zhang, and M. Chen, "Smart sensing in building and construction: Current status and future perspectives," *Sensors*, vol. 24, no. 7, p. 2156, 2024.

[38] J. Wang, X. Liu, and H. Chen, "Emerging iot technologies for smart environments: Applications and challenges," *Sensors*, vol. 24, no. 6, p. 1892, 2024.

[39] K. Zhang, Y. Li, Z. Wang, and X. Zhao, "Cyber-physical systems security: A systematic review," *Computers & Industrial Engineering*, vol. 165, p. 107914, 2022.

[40] R. Yohanandhan, R. Elavarasan, P. Manoharan, and L. Mihet-Popa, "Cyber-physical systems and smart cities in india: Opportunities, issues, and challenges," *Sensors*, vol. 21, no. 22, p. 7714, 2021.

[41] J. Kapusta, W. Bauer, and J. Baranowski, "Evaluation of the effectiveness of physical protection systems with consideration of its cyber-resilience," in *2023 27th International Conference on Methods and Models in Automation and Robotics (MMAR)*, pp. 457–461, IEEE, 2023.

[42] J. Kapusta, W. Bauer, and J. Baranowski, "Detection of building infrastructure elements in layouts using yolo as a foundation for graph-based adversarial attack path modeling," in *ICON 2024: International Conference on Security Technologies and Applications*, 2025. To appear.

[43] J. Kapusta, W. Bauer, and J. Baranowski, "A heuristic-based framework for risk evaluation of attack paths in cyber-physical protection systems," in *2025 29th International Conference on Methods and Models in Automation and Robotics (MMAR)*, (Międzyzdroje, Poland), IEEE, 2025. To appear.

[44] A. Author, "A brief title," *Journal Name*, vol. 1, pp. 1–10, 1981.

[45] I. A. E. Agency, "Development, use and maintenance of the design basis threat: Implementing guide," Tech. Rep. IAEA Nuclear Security Series No. 10, IAEA, Vienna, 2009.

[46] M. AlMuhaini, "Resilience quantification model for cyber-physical power systems," *IEEE Transactions on Power Systems*, 2024.

[47] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.

[48] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[49] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.

[50] G. Jocher, A. Chaurasia, and J. Qiu, "Yolo by ultralytics." `https://github.com/ultralytics/ultralytics`, 2023. Accessed: 2024-06-01.

[51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, pp. 8024–8035, 2019.

[52] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[53] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, IEEE, 2017.

[54] R. J. Wilson, *Introduction to Graph Theory*. Pearson, 5th ed., 2010.

[55] R. Diestel, *Graph Theory*. Springer Graduate Texts in Mathematics, 5th ed., 2017.

[56] R. Lindelauf, P. Borm, and H. Hamers, "The influence of secrecy on the communication structure of covert networks," *Social Networks*, vol. 31, no. 2, pp. 126–137, 2009.

[57] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, "Network analysis in the social sciences," *Science*, vol. 323, no. 5916, pp. 892–895, 2009.

[58] M. Newman, *Networks: An Introduction*. Oxford University Press, 2010.

[59] R. S. Garfinkel and G. L. Nemhauser, *Integer Programming*. John Wiley & Sons, 1972.

[60] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993.

[61] T. Alpcan and T. Başar, *Network Security: A Decision and Game-Theoretic Approach*. Cambridge University Press, 2011.

[62] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Başar, and J.-P. Hubaux, "Game theory meets network security and privacy," *ACM Computing Surveys*, vol. 45, no. 3, pp. 1–39, 2013.

[63] T. L. Norman, *Risk Analysis and Security Countermeasure Selection*. CRC Press, 2010.

[64] L. J. Fennelly, *Effective Physical Security*. Butterworth-Heinemann, 4th ed., 2013.

[65] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[66] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?," *Brain Informatics*, vol. 3, no. 2, pp. 119–131, 2016.

[67] Committee of Sponsoring Organizations of the Treadway Commission, *Enterprise Risk Management – Integrating with Strategy and Performance*. Durham, NC: American Institute of CPAs, 2017. Executive Summary available at https://www.coso.org/Pages/erm.aspx.

[68] J. Bang-Jensen and G. Gutin, *Digraphs: Theory, Algorithms and Applications*. Springer Science & Business Media, 2009.

[69] R. Tarjan, "Depth-first search and linear graph algorithms," *SIAM Journal on Computing*, vol. 1, no. 2, pp. 146–160, 1972.

[70] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. MIT Press, 3rd ed., 2009.

[71] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[72] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[73] D. B. Johnson, "Finding all the elementary circuits of a directed graph," *SIAM Journal on Computing*, vol. 4, no. 1, pp. 77–84, 1975.

[74] R. Sedgewick and K. Wayne, *Algorithms*. Addison-Wesley Professional, 4th ed., 2011.

[75] S. S. Skiena, *The Algorithm Design Manual*. Springer, 2nd ed., 2008.

[76] D. E. Knuth, *The Art of Computer Programming, Volume 1: Fundamental Algorithms*. Addison-Wesley, 3rd ed., 1997.

[77] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *Data Structures and Algorithms*. Addison-Wesley, 1983.

[78] V. Batagelj and A. Mrvar, "A subquadratic triad census algorithm for large sparse networks with small maximum degree," *Social Networks*, vol. 23, no. 3, pp. 237–243, 2001.

[79] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*. Cambridge University Press, 2nd ed., 2014.

[80] E. F. Moore, "The shortest path through a maze," *Proceedings of the International Symposium on the Theory of Switching*, pp. 285–292, 1959.

[81] C. Y. Lee, "An algorithm for path connections and its applications," *IRE Transactions on Electronic Computers*, vol. EC-10, no. 3, pp. 346–365, 1961.

[82] J. Hopcroft and R. Tarjan, "Algorithm 447: efficient algorithms for graph manipulation," *Communications of the ACM*, vol. 16, no. 6, pp. 372–378, 1973.

[83] S. R. Kosaraju, *Algorithms for some graph problems*. PhD thesis, Johns Hopkins University, 1972.

[84] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[85] M. L. Fredman and R. E. Tarjan, "Fibonacci heaps and their uses in improved network optimization algorithms," *Journal of the ACM*, vol. 34, no. 3, pp. 596–615, 1987.

[86] D. B. Johnson, "Efficient algorithms for shortest paths in sparse networks," *Journal of the ACM*, vol. 24, no. 1, pp. 1–13, 1977.

[87] I. Pohl, "Heuristic search viewed as path finding in a graph," *Artificial Intelligence*, vol. 1, no. 3, pp. 193–204, 1970.

[88] R. E. Korf, "Real-time heuristic search," *Artificial Intelligence*, vol. 42, no. 2-3, pp. 189–211, 1990.

[89] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[90] J. Pearl, *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley, 1984.

[91] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Pearson, 4th ed., 2020.

[92] N. J. Nilsson, *Principles of Artificial Intelligence*. Morgan Kaufmann, 1980.

[93] R. Dechter and J. Pearl, "Generalized best-first search strategies and the optimality of a*," *Journal of the ACM*, vol. 32, no. 3, pp. 505–536, 1985.

[94] A. Martelli, "On the complexity of admissible search algorithms," *Artificial Intelligence*, vol. 8, no. 1, pp. 1–13, 1977.

[95] J. Pearl, "Heuristics and the consistency condition," *Artificial Intelligence*, vol. 10, no. 2, pp. 219–223, 1978.

[96] L. Mero, "A heuristic search algorithm with modifiable estimate," *Artificial Intelligence*, vol. 23, no. 1, pp. 13–27, 1984.

[97] A. Bagchi and A. Mahanti, "Search algorithms under different kinds of heuristics—a comparative study," *Journal of the ACM*, vol. 30, no. 1, pp. 1–21, 1983.

[98] R. E. Korf, M. Reid, and S. Edelkamp, "Time complexity of iterative-deepening-a*," *Artificial Intelligence*, vol. 129, no. 1-2, pp. 199–218, 2001.

[99] I. Pohl, "The avoidance of (relative) catastrophe, heuristic competence, genuine dynamic weighting and computational issues in heuristic problem solving," in *Proceedings of the 3rd International Joint Conference on Artificial Intelligence*, pp. 12–17, 1973.

[100] E. A. Hansen and R. Zhou, "Anytime heuristic search," *Journal of Artificial Intelligence Research*, vol. 28, pp. 267–297, 2007.

[101] M. Likhachev, G. J. Gordon, and S. Thrun, "Ara*: Anytime a* with provable bounds on sub-optimality," in *Advances in Neural Information Processing Systems*, vol. 16, 2003.

[102] H. Kaindl and G. Kainz, "Bidirectional heuristic search reconsidered," *Journal of Artificial Intelligence Research*, vol. 7, pp. 283–317, 1997.

[103] R. C. Holte, A. Felner, G. Sharon, and N. R. Sturtevant, "Bidirectional search that is guaranteed to meet in the middle," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pp. 3411–3417, 2016.

[104] F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*. Springer-Verlag, 1985.

[105] A. Author, "Computational title," *Journal Name*, vol. 1, pp. 1–10, 2008.

[106] A. Rosenfeld and J. L. Pfaltz, "Distance functions on digital pictures," *Pattern Recognition*, vol. 1, no. 1, pp. 33–61, 1968.

[107] B. Hillier and J. Hanson, *The Social Logic of Space*. Cambridge University Press, 1984.

[108] A. Turner, "From axial to road-centre lines: a new representation for space syntax and a new model of route choice for transport network analysis," *Environment and Planning B*, vol. 34, no. 3, pp. 539–555, 2007.

[109] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[110] J. Nash, "Non-cooperative games," *The Annals of Mathematics*, vol. 54, no. 2, pp. 286–295, 1951.

[111] J. C. Harsanyi, "Games with incomplete information played by bayesian players, i-iii," *Management Science*, vol. 14, no. 3, pp. 159–182, 1967.

[112] D. Fudenberg and J. Tirole, *Game Theory*. MIT Press, 1991.

[113] E. Von Stack, "Market analysis in 1934," *Economic Journal*, vol. 1, pp. 1–20, 1934.

[114] A. Ben, "Robust something," *Journal Name*, vol. 1, pp. 1–10, 2009.

[115] D. Bertsimas, D. B. Brown, and C. Caramanis, "Theory and applications of robust optimization," *SIAM Review*, vol. 53, no. 3, pp. 464–501, 2011.

[116] P. Kouvelis and G. Yu, *Robust Discrete Optimization and Its Applications*. Kluwer Academic Publishers, 1997.

[117] J. M. Mulvey, R. J. Vanderbei, and S. A. Zenios, "Robust optimization of large-scale systems," *Operations Research*, vol. 43, no. 2, pp. 264–281, 1995.

[118] N. Guarino, "Formal ontology and information systems," in *Proceedings of the First International Conference on Formal Ontology in Information Systems*, pp. 3–15, 1998.

[119] B. Smith, "Ontology and information systems," in *Handbook on Ontologies* (S. Staab and R. Studer, eds.), pp. 15–34, Berlin/Heidelberg, Germany: Springer, 2004.

[120] G. Frege, "Über sinn und bedeutung," *Zeitschrift für Philosophie und philosophische Kritik*, vol. 100, pp. 25–50, 1892.

[121] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, 2015.

[122] V. Pareto, *Cours d'économie politique*. Rouge, 1896.

[123] M. Ehrgott, *Multicriteria Optimization*. Springer, 2nd ed., 2005.

[124] K. Miettinen, *Nonlinear Multiobjective Optimization*. Kluwer Academic Publishers, 1999.

[125] J. Y. Yen, "Finding the k shortest loopless paths in a network," *Management Science*, vol. 17, no. 11, pp. 712–716, 1971.

[126] D. Eppstein, "Finding the k shortest paths," *SIAM Journal on Computing*, vol. 28, no. 2, pp. 652–673, 1998.

[127] E. Q. Martins and M. M. Pascoal, "A new implementation of yen's ranking loopless paths algorithm," *4OR*, vol. 1, no. 2, pp. 121–133, 2003.

[128] L. Ford and D. Fulkerson, *Flows in Networks*. Princeton University Press, 1962.

[129] T. Chondrogiannis, P. Bouros, J. Gamper, and U. Leser, "Alternative routing: k-shortest paths with limited overlap," in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 1–4, 2015.

[130] E. Tsang, *Foundations of Constraint Satisfaction*. Academic Press, 1993.

[131] R. Dechter, *Constraint Processing*. Morgan Kaufmann, 2003.

[132] A. K. Mackworth, "Consistency in networks of relations," *Artificial Intelligence*, vol. 8, no. 1, pp. 99–118, 1977.

[133] J. O'Rourke, *Computational Geometry in C*. Cambridge University Press, 2nd ed., 1998.

[134] J. L. Bentley and T. A. Ottmann, "Algorithms for reporting and counting geometric intersections," *IEEE Transactions on Computers*, vol. C-28, no. 9, pp. 643–647, 1979.

[135] B. Delaunay, "Sur la sphère vide," *Bulletin de l'Académie des Sciences de l'URSS*, vol. 6, pp. 793–800, 1934.

[136] R. E. Barlow and F. Proschan, *Mathematical Theory of Reliability*. John Wiley & Sons, 1965.

[137] M. Rausand and A. Høyland, *System Reliability Theory: Models, Statistical Methods, and Applications*. John Wiley & Sons, 2nd ed., 2004.

[138] S. M. Ross, *Introduction to Probability Models*. Academic Press, 11th ed., 2014.

[139] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 4th ed., 2002.

[140] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd ed., 2013.

[141] C. P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, 2nd ed., 2007.

[142] E. A. Lee, "Cyber physical systems: Design challenges," in *Proceedings of the 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing*, pp. 363–369, 2008.

[143] R. Rajkumar, I. Lee, L. Sha, and J. Stankovic, "Cyber-physical systems: the next computing revolution," in *Proceedings of the 47th Design Automation Conference*, pp. 731–736, 2010.

[144] R. Alur, C. Courcoubetis, N. Halbwachs, T. A. Henzinger, P.-H. Ho, X. Nicollin, A. Olivero, J. Sifakis, and S. Yovine, "The algorithmic analysis of hybrid systems," *Theoretical Computer Science*, vol. 138, no. 1, pp. 3–34, 1995.

[145] A. Bar, "Dynamics of something," *Journal Name*, vol. 1, pp. 1–10, 1997.

[146] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[147] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[148] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, "Catastrophic cascade of failures in interdependent networks," *Nature*, vol. 464, no. 7291, pp. 1025–1028, 2010.

[149] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[150] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[151] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Brussels, Belgium), pp. 66–71, Association for Computational Linguistics, 2018.

[152] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, *et al.*, "Emergent abilities of large language models," *Transactions on Machine Learning Research*, 2022.

[153] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837, 2022.

[154] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[155] B. T. Willard and R. Louf, "Efficient guided generation for large language models," *arXiv preprint arXiv:2307.09702*, 2023.

[156] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[157] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[158] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 1999.

[159] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[160] International Organization for Standardization, Geneva, Switzerland, *Risk management – Guidelines*, second edition ed., 2018. Corrected version 2018-08.

[161] L. J. Fennelly, *Effective Physical Security*. Oxford: Butterworth-Heinemann, 4th ed., 2013.

[162] M. L. Garcia, *The Design and Evaluation of Physical Protection Systems*. Oxford: Butterworth-Heinemann, 2nd ed., 2008.

[163] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?," *Brain Informatics*, vol. 3, no. 2, pp. 119–131, 2016.

[164] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[165] T. L. Norman, *Risk Analysis and Security Countermeasure Selection*. Boca Raton, FL: CRC Press, 2010.