UNIVERZITA PAVLA JOZEFA ŠAFÁRIKA V KOŠICIACH

PRÍRODOVEDECKÁ FAKULTA

IDENTIFIKÁCIA VZŤAHOV A RELEVANTNÝCH STÔP V RÁMCI DIGITÁLNEJ FORENZNEJ ANALÝZY

2023 **Mgr. Eva MARKOVÁ**

UNIVERZITA PAVLA JOZEFA ŠAFÁRIKA V KOŠICIACH

PRÍRODOVEDECKÁ FAKULTA

IDENTIFIKÁCIA VZŤAHOV A RELEVANTNÝCH STÔP V RÁMCI DIGITÁLNEJ FORENZNEJ ANALÝZY

RIGORÓZNA PRÁCA

Študijný program:	Informatika
Pracovisko (katedra/ústav):	Ústav informatiky
Vedúci záverečnej práce:	doc. RNDr. JUDr. Pavol Sokol, PhD.

Košice 2023 **Mgr. Eva MARKOVÁ**

Poďakovanie

Týmto sa chcem poďakovať vedúcemu svojej práce doc. RNDr. JUDr. Pavlovi Sokolovi, PhD. za odborné a profesionálne vedenie, cenné rady a veľkú pomoc počas tvorby práce.

Abstrakt v štátnom jazyku

Analýza digitálnych stôp je kľúčovou pri riešení bezpečnostných incidentov. Samotná digitálna forenzná analýza môže vzhľadom k typu a závažnosti bezpečnostného incidentu zabrať množstvo času, preto je dôležité niektoré fázy digitálnej forenznej analýzy čiastočne automatizovať. V práci sa venujeme analýze digitálnych stôp pomocou metód strojového učenia, konkrétne so zameraním na metódy bez učiteľa (ECOD, LOF, PCA a IForest). Takisto vysvetľujeme základnú terminológiu a prezentujeme použitý dataset na skúmanie týchto metód. Tiež poskytujeme parciálne výsledky. Vo väčšine prípadov behov metód sme najlepšie výsledky dosiahli pomocou bezparametrickej metódy ECOD, ktorá bola aj najrýchlejšia.

Kľúčové slová: digitálna stopa, digitálna forenzná analýza, ECOD, IForest, LOF, PCA

Abstrakt v cudzom jazyku

Analysis of digital evidence is key to solving security incidents. The digital forensics itself can take a lot of time due to the type and severity of the security incident, so it is important to partially automate some phases of the digital forensics. In our thesis, we are devoted to the analysis of digital evidence using machine learning methods, specifically with a focus on unsupervised methods (ECOD, LOF, PCA and IForest). We also explain the basic terminology and present the dataset used to investigate these methods. We also provide partial results. In most cases of method runs, we achieved the best results using the non-parametric ECOD method, which was also the fastest.

Keywords: digital evidence, digital forensics, ECOD, IForest, LOF, PCA

Obsah

Zoznam ilustrácií

Zoznam tabuliek

Zoznam skratiek a značiek

AUPR	Area Under the Precision-Recall curve
AUC	Area Under the ROC curve
AVPR	Average Precision
ARP	Address Resolution Protocol
CTF	Capture The Flag
CDF	Cumulative Distribution Function
ECOD	Empirical Cumulative Distribution-based Outlier Detection
FCA	Formal Concept Analysis
IForest	Isolation Forest
LOF	Local Outlier Factor
MCC	Matthews Correlation Coefficient
MFT	Master File Table
PCA	Principal Component Analysis
RFC	Request For Comments
ROC	Receiver Operating Characteristic
SWGDE	Scientific Working Group on Digital Evidence

Úvod

V súčasnej dobe sa informačná a kybernetická bezpečnosť stretáva s významným vzostupom kybernetických hrozieb, ktoré sa neustále vyvíjajú a stávajú sa čoraz sofistikovanejšími. Táto situácia má za následok rýchlejšie a presnejšie uskutočňovanie útokov. Zároveň sledujeme posun od jednoduchých a masových útokov k cieleným, ktoré smerujú priamo na konkrétne organizácie a inštitúcie. V tomto kontexte je pre organizácie nevyhnutné vyvinúť schopnosť reagovať na bezpečnostné hrozby a riešiť bezpečnostné incidenty.

V tejto dynamicky sa meniacej sa oblasti zohráva kľúčovú úlohu digitálna forenzná analýza. Digitálni forenzní analytici potrebujú rýchly prehľad o aktuálnom stave udalostí a prístup k relevantným informáciám týkajúcim sa daného vyšetrovaného prípadu. Cieľom pri riešení bezpečnostných incidentov je identifikovať príčiny týchto udalostí, analyzovať postup útočníka a jeho cieľ, aby sa organizácia mohla z týchto skúseností poučiť a zabezpečiť sa do budúcnosti. Obzvlášť pri komplexných bezpečnostných incidentoch je nevyhnutné sa zamerať na analýzu veľkého množstva digitálnych stôp, čo predstavuje rôzne výzvy. Pri vyšetrovaných prípadoch sa digitálni forenzní analytici musia rozhodnúť, kde začať svoju analýzu a akú digitálnu stopu sledovať. V kontexte narastajúcej komplexity a stôp dát je kľúčové vytvoriť systematický a efektívny postup pre správu digitálnych stôp a stanovenie forenzných hypotéz, ktoré následne musia byť overené.

V množstve digitálnych stôp získaných zo rôznych zariadení nie sú všetky stopy nevyhnutné a relevantné pre dané vyšetrovanie. Z toho dôvodu je kritická schopnosť oddeľovať digitálne stopy, ktoré majú význam pre analyzovaný prípad, od tých, ktoré nemajú. Väčšina existujúcich techník vyžaduje manuálne vyhľadávanie a selekciu, čo môže byť časovo náročné a náchylné na ľudské chyby [1]. Na druhej strane, prístup s využitím techník strojového učenia s učiteľom a bez učiteľa, má potenciál výrazne urýchliť a zefektívniť digitálny forenzný proces. Tieto metódy môžu automaticky identifikovať vzory a anomálie v digitálnych stopách, čo zjednodušuje detekciu relevantných stôp. Týmto spôsobom je možné efektívne zúžiť priestor pre ďalšiu manuálnu analýzu iba na dôležité a potenciálne relevantné dáta.

V tejto práci sa zamýšľame nad touto problematikou a zameriavame sa na možnosť automatickej identifikácie neobvyklých výskytov digitálnych stôp, ktoré je

možné získať zo súborového systému. Väčšina známych prístupov sa zameriava na hľadanie vzorcov; majú určité znalosti na vstupe – my pracujeme „bez znalostí“ – na vstupe nedostávame nič, čo by mohlo naznačovať, že záznam je anomália alebo relevantná digitálna stopa.

Hlavným cieľom tejto práce je analyzovať možnosti použitia metód strojového učenia pri fáze analýzy digitálnych stôp vzhľadom na komplexnosť, množstvo a heterogénnosť forenzných artefaktov. Tento cieľ môžeme rozdeliť na tri hlavné podciele a to yytvorenie vhodného datasetu pre porovnávanie metód pri analýze digitálnych stôp, identifikácia relevantných digitálnych stôp a hľadanie vzťahov medzi digitálnymi stopami a ich atribútmi. Dôležitým aspektom je tiež nájdenie vhodného spôsobu, ako vyhodnocovať výsledky použitých metód strojového učenia.

Práca sa skladá zo štyroch hlavných kapitol. V prvej kapitole sa venujeme základným pojmom ako digitálna forenzná analýza, digitálna stopa a fázy digitálnej forenznej analýzy. Popisujeme, aké činnosti zahŕňa každá z fáz digitálnej forenznej analýzy a takisto opisujeme výzvy v rámci tejto oblasti vzhľadom k využitiu metód strojového učenia. V druhej kapitole sa bližšie venujeme prehľadu aktuálneho stavu riešenia danej problematiky a takisto nášmu prístupu riešenia pre každý výskumný cieľ (vytvorenie datasetu, identifikácia relevantných digitálnych stôp a hľadanie vzťahov medzi relevantnými digitálnymi stopami). V tretej kapitole popisujeme ukážku prístupu riešenia. Pre pochopenie kontextu popisujeme prípad, dataset a predspracovanie datasetu. Tiež sa venujeme použitým metódam strojového učenia bez učiteľa a metrikám, ktoré využívame na vyhodnocovanie pri evaluácii. Posledná kapitola je venovaná parciálnym výsledkom z nášho výskumu. Uvádzame celkové výsledky, analyzujeme výkon štyroch vybraných metód. Sledujeme ako vplýva výber atribútov, agregačných funkcií a hodnôt rôznych parametrov na celkové výsledky. Na záver sme analyzovali samotných outlierov (inody súborov aj názvy súborov).

1 Digitálna forenzná analýza

Forenzná veda môže byť definovaná ako aplikácia vedeckých princípov na právne účely. Digitálna forenzná analýza je veľmi dôležitým komponentom samotnej odpovede na bezpečnostný incident, čo je porušenie alebo bezprostredná hrozba porušenia pravidiel počítačovej bezpečnosti, prijateľných zásad používania alebo štandardných bezpečnostných postupov. Jedná sa o aplikáciu metód digitálnej forenznej analýzy na pochopenie sledu udalostí v systéme, ktoré by mohli viesť ku škodlivej aktivite. Digitálna forenzná analýza je oblasť forenznej vedy, ktorá sa zameriava na identifikáciu, získavanie, spracovanie, analýzu a podávanie správ o údajoch uložených elektronicky [2].

1.1 Digitálna stopa

Fyzické zariadenia ako napríklad počítačové systémy, mobilné telefóny, USB kľúče, pamäťové karty, smerovače, prepínače, modemy a podobne môžu obsahovať digitálne stopy. Vo všeobecnosti zdrojmi digitálnych stôp môžu byť otvorené počítačové systémy (disky, pamäť, prenosné médiá a podobne), komunikačné systémy a embedované zariadenia. Podľa SWGDE [3], digitálna stopa je akákoľvek informácia s vypovedacou hodnotou uložená alebo prenášaná v digitálnej binárnej forme, ktorá môže byť predložená súdu ako vecný dôkaz s vypovedacou hodnotou. Vo všeobecnosti však vysoký stupeň ochrany dát môže sťažovať alebo znemožniť prácu s digitálnymi stopami [4].

1.2 Fázy digitálnej forenznej analýzy

Existuje niekoľko rôznych schém, ktoré popisujú samotný proces digitálnej forenznej analýzy, vo väčšine prípadov je však výsledok rovnaký. V našom prípade uvažujeme „Digital Investigate Framework“, ktorý bol vytvorený v rámci „Digital Forensics Research Workshop“ (DFRWS). Tento rámec nám poskytuje 6 krokov digitálnej forenznej analýzy [5]:

1. Identifikácia
2. Uchovanie
3. Zber

4. Vyťažovanie

5. Analýza

6. Prezentácia

Cieľom digitálnej forenznej analýzy je zozbierať všetky relevantné digitálne stopy, zaistiť ich a spracovať ich. Dĺžka vykonávania sa môže líšiť v závislosti od konkrétneho incidentu a jeho typu. V ďalších podkapitolách popíšeme jednotlivé fázy procesu digitálnej forenznej analýzy.

1.2.1 Identifikácia

Jedným z často diskutovaných princípov v rámci digitálnej forenznej analýzy je Locardov princíp výmeny. Hovorí o tom, že ak dva objekty prídu do kontaktu, zanechajú na sebe navzájom stopy. Tento princíp sa dá preniesť aj do digitálneho sveta – takisto ak dva systémy spolu komunikujú, v oboch systémoch ostávajú stopy (za predpokladu, že sa útočník nesnaží odstrániť akékoľvek zmienky v systéme o svojej činnosti). Cieľom tejto fázy je identifikácia potenciálnych zdrojov digitálnych stôp vytvorených v čase incidentu [5].

1.2.2 Uchovanie

Bezprostredne po identifikácii je potrebné zabrániť akejkoľvek modifikácii alebo vymazaniu digitálnych stôp. Takisto je potrebné zabezpečiť, aby sa akýkoľvek neoprávnený používateľ nedostal k zaisteným zariadeniam. Pri virtuálnych platformách sa neoprávnenému zaobchádzaniu predchádza vytváraním tzv. „snapshotov“ systému a ukladaním na trvalé úložiská [4].

1.2.3 Zber a zaisťovanie

Pri zbere a zaisťovaní je potrebné brať do úvahy volatilitu stôp. Niektoré stopy totižto môžu byť zničené napríklad po vypnutí zariadenia. RFC 3227 nám poskytuje poradie volatility digitálnych stôp [6]:

1. Registre, pamäť cache

2. smerovacia tabuľka, ARP cache, tabuľka procesov, údaje o kerneli, pamäť (RAM)

3. Dočasné súborové systémy

4. Disk

5. Dáta o vzdialenom monitoringu a logovaní

6. Fyzické konfigurácie, sieťová topológia

7. Archívne médiá

Chyby pri zbere a zaisťovaní môžu viesť k poškodeniu a nespoľahlivosti digitálnych stôp. Činnosti vykonávané analytikmi by nemali meniť pôvodné stopy a takisto je potrebné, aby analytici všetko dokumentovali. Na dokumentáciu pri zaisťovaní slúži tzv. „Chain of custody“, čo je dokument slúžiaci na dokumentáciu zaistených stôp počas ich životného cyklu (dokumentácia sa končí buď vrátením, alebo zničením digitálnej stopy) [5].

1.2.4 Vyťažovanie

Táto fáza je o výbere konkrétnych nástrojov a forenzných techník slúžiacich na objavenie a extrakciu údajov zo stôp, ktoré súvisia s incidentom. Skúmanie digitálnych stôp tiež pokračuje v procese uchovávania tým, že analytici počas skúmania zachovávajú maximálnu starostlivosť o stopy. V prípade, že sa forenzný analytik v tejto fáze nepostará o uchovanie dôkazov, existuje možnosť kontaminácie, ktorá by viedla k tomu, že stopy budú nespoľahlivé alebo nepoužiteľné [5].

1.2.5 Analýza

Akonáhle sa vo fáze vyťažovania extrahujú potenciálne relevantné časti údajov, forenzný analytik analyzuje údaje v korelácii s inými získanými relevantnými údajmi. Cieľom analýzy je potvrdiť alebo vyvrátiť hypotézy súvisiace s vyšetrovaním. Počas analýzy je analytik povinný zachovať integritu digitálnych stôp [4].

1.2.6 Prezentácia

Reportovanie faktov súvisiacich s digitálnou forenznou analýzou musí byť jasné, stručné a so zachovaním objektivity. Takmer vo všetkých prípadoch sa od forenzného analytika vyžaduje, aby pripravil podrobnú písomnú správu, ktorá sa zaoberá každým úkonom a zachytí potrebné kritické údaje. Správa by mala byť dôkladná, presná a bez názorov alebo zaujatosti [5].

1.3 Výzvy digitálnej forenznej analýzy

Pri realizácii digitálnej forenznej analýzy sa naskytuje perspektíva využitia metodológií strojového učenia. Strojové učenie označuje súbor algoritmov, ktoré získavajú poznatky z dát a na základe týchto poznatkov vykonávajú rozhodnutia. V oblasti digitálnej forenznej analýzy existuje viacero procesov, ktoré by mohli byť potenciálne riešené prostredníctvom metód strojového učenia, avšak doteraz neboli tieto riešenia plne implementované.

Využitie metód strojového učenia prináša so sebou rôzne náročné úlohy, ktoré je dôležité skúmať a zdokonaľovať. Tieto techniky majú veľký potenciál pri spracovaní a korelácii údajov z viacerých zdrojov. Napríklad vo vedeckej práci [7] autori zhromažďujú varovania z rôznych zdrojov, normalizujú ich a následne pomocou strojového učenia kategorizujú. V ďalšom výskume [8] sa využívajú metódy strojového učenia na koreláciu udalostí s cieľom detekcie anomálií. Táto práca analyzuje a spája udalosti za účelom odhalenia neobvyklých a potenciálne škodlivých vzorov.

Jednou z oblastí, kde je možné využiť strojové učenie je triedenie (triáž) forenzných stôp. Pre triáž bola v roku 2013 navrhnutá metóda založená na strojovom učení [9]. Cieľom tohto navrhnutého prístupu bolo automatizovať kategorizáciu digitálnych médií na základe identifikovaných možných vzťahov medzi získanými stopami a skúmanými trestnými činmi. V priebehu svojho výskumu autori preskúmali rôzne techniky, ako napríklad Bayesovské siete, rozhodovacie stromy, metódu podporných vektorov a iné, s cieľom identifikovať tie, ktoré najlepšie zapadajú do konkrétneho scenára.

Inou oblasťou skúmania sú metadáta súborových systémov [10, 11]. Tieto údaje zaznamenávajú a uchovávajú posledné akcie so súbormi, vrátane dátumu vytvorenia,

posledného prístupu a úpravy. Digitálne vyšetrovanie sa sústreďuje na získavanie relevantných informácií z týchto metadát a časovej osi, aby sa identifikovali položky s významnou forenznou hodnotou. Metadáta, ako napríklad veľkosť súboru, cesta k súboru či názov súboru, sú často využívané na filtrovanie a indexovanie súborov.

Metadáta sú úzko prepojené s problematikou vytvárania a analýzy časových osí udalostí, kde časová os zohráva významnú úlohu pri získavaní celkového prehľadu o skúmanom prípade. V súčasnosti sa na tieto účely využíva nástroj Log2timeline (plaso) [12], ktorý umožňuje generovať tzv. "super časové osi". Tento nástroj zohľadňuje digitálne udalosti zo súborového systému, registrov, sieťových protokolov a aplikačných protokolov [13]. V článkoch [14, 15] sa autori zaoberajú efektívnym využitím najmodernejších techník strojového učenia a analýzy sentimentu v rámci forenznej časovej osi.

Jednou z významných výziev v oblasti digitálnej forenznej analýzy je nedostatok kvalitných datasetov [7, 16, 17]. Dôležité je vytvoriť dataset, na ktorom by sa dali otestovať metódy strojového učenia. Tejto problematike sa venujeme v kapitole 2.1.

Vo všeobecnosti teda existuje mnoho výziev v oblasti digitálnej forenznej analýzy, v rámci ktorých je možné použiť metódy strojového učenia. Dôležité je si stanoviť podoblasť, ktorou sa chceme zaoberať. V našom výskume riešime práve nedostatok datasetov, identifikáciu relevantných digitálnych stôp a tiež hľadanie vzťahov medzi nimi.

2 Výskumné ciele

V predchádzajúcej kapitole sme načrtli niekoľko aktuálnych výziev súčasnej digitálnej forenznej analýzy. Ako sme už viackrát načrtli v tejto práci, digitálna forenzná analýza je komplexná záležitosť, ktorej cieľom je zistiť, čo sa reálne stalo a týmto pomôcť najmä pri reakcii na bezpečnostné incidenty a riešení kybernetickej kriminality (vzhľadom na rozsah práce sa primárne zameriavame na reakciu na bezpečnostné incidenty). Jeden z hlavných problémov je hľadanie toho, čo je relevantné pre samotnú analýzu (vyhodnotenie konkrétnej stopy) a následne hľadanie vzťahov medzi týmito stopami. Z týchto dôvodov hlavným cieľom tejto práce je analyzovať možnosti použitia metód strojového učenia pri fáze analýzy digitálnych stôp vzhľadom na komplexnosť, množstvo a heterogénnosť forenzných artefaktov. Forenzné artefakty sú objekty, ktoré majú forenznú hodnotu, a teda obsahujú údaje alebo stopy niečoho, čo sa udialo v systéme.

Identifikácia relevantných digitálnych stôp predstavuje základný problém pri riešení bezpečnostných incidentov. V rámci rýchleho a adekvátneho riešenia bezpečnostných incidentov je nevyhnutné identifikovať zdroje digitálnych stôp, zaistiť digitálne stopy a extrahovať digitálne stopy relevantné pre konkrétny prípad, resp. skúmanú otázku. Na tomto mieste sa otvára otázka, akým spôsobom by bolo možné tento proces automatizovať. Vhodný spôsob riešenia nám poskytujú rôzne metódy strojového učenia. Na základe vyššie uvedeného sme si stanovili nasledujúci výskumný cieľ:

- **Identifikácia relevantných digitálnych stôp pomocou metód strojového učenia a vyhodnotenie efektívnosti týchto metód.**

Ako sme už vyššie uviedli, v rámci reakcie na bezpečnostný incident a pre forenzné vyšetrovanie nie je dôležité len identifikovať relevantné digitálne stopy, ale súčasne aj hľadať vzťahy medzi týmito stopami. Tieto vzťahy nám poskytujú ďalšie dôležité informácie, najmä z pohľadu určitého kontextu. Dokážu odpovedať na otázky týkajúce sa časového poradia konkrétnych udalostí, relevantnosti konkrétnych digitálnych stôp pre forenzné vyšetrovanie, identifikácie konkrétnych fáz útoku podľa rôznych bezpečnostných rámcov (napr. MITRE ATTACK). Na tomto mieste je vhodné uvažovať nad tým, ako je možné pomôcť samotnému analytikovi a nazerať sa na daný

problém komplexnejšie, ako by to vykonal samotný analytik. Na základe vyššie uvedeného sme si stanovili nasledujúci výskumný cieľ:

- **Hľadanie vzťahov medzi relevantnými digitálnymi stopami pomocou metód strojového učenia.**

Vyššie uvedené výskumné ciele vyžadujú okrem stanovenia konkrétnej otázky a použitia vhodných metód aj použitie vhodnej dátovej sady (datasetu). V súčasnej dobe sa stretávame s viacerými problémami súvisiacimi s použitím vhodného datasetu. Prvým problémom je nemožnosť použitia dát z reálneho prípadu vzhľadom na citlivosť týchto údajov a rôzne bezpečnostné a právne prekážky. Druhý problém predstavuje komplexnosť syntetických datasetov, ktoré viackrát nepokrývajú skutočné situácie. Na základe vyššie uvedeného sme si stanovili nasledujúci výskumný cieľ:

- **Vytvorenie vhodného datasetu pre identifikáciu relevantných digitálnych stôp a hľadanie vzťahov medzi nimi.**

Zaujímavou oblasťou sú aj anti-forenzné techniky, ktoré sú schopné narušiť a predĺžiť proces digitálneho forenzného vyšetrovania a dotýkajú sa každého z vyššie uvedených cieľov. Otázkami, ktorými je potrebné sa pri používaní anti-forenzných techník zaoberať je viacero. Pre lepšie porozumenie týchto metód je nevyhnutné skúmať, ako ovplyvňujú proces identifikácie a rekonštrukcie digitálnych stôp. Do budúcna chceme preskúmať, ako možno vytvoriť reprezentatívny dataset, ktorý zohľadňuje rôznorodé anti-forenzné techniky. Takýto dataset je nevyhnutný na testovanie efektívnosti analytických nástrojov a techník pri identifikácii stôp, ktoré boli úmyselne zamaskované alebo zmenené pomocou anti-forenzných techník. Ďalej chceme skúmať, akým spôsobom anti-forenzné techniky ovplyvňujú schopnosť identifikovať relevantné digitálne stopy. Tieto techniky môžu zahŕňať napríklad modifikáciu metadát, úpravu časových pečiatok alebo použitie šifrovania na skrytie obsahu. Dôsledkom týchto opatrení môže byť zníženie schopnosti forenzných analytikov správne identifikovať zdroj, pôvod a čas vytvorenia digitálnych stôp. Posledným dôležitým bodom je preskúmanie vplyvu anti-forenzných techník na hľadanie vzťahov medzi rôznymi digitálnymi stopami. Digitálne stopy často vytvárajú spojenia medzi rôznymi udalosťami

a osobami, avšak anti-forenzné metódy môžu tieto vzťahy zamaskovať, čo značne komplikuje prácu analytikov pri rekonštrukcii udalostí a zisťovaní kontextu.

Nasledujúca časť kapitoly je rozdelená na podkapitoly, v rámci ktorých podrobnejšie popisujeme prehľad súčasného stavu riešenej problematiky a tiež náš prístup k jej riešeniu pre každý výskumný cieľ.

2.1 Vytvorenie vhodného datasetu

Významným aspektom pri výskume nad údajmi súvisiacimi s digitálnou forenznou analýzou je schopnosť vytvoriť dataset, ktorý by spĺňal isté očakávania a požiadavky. Vo všeobecnosti totižto neexistuje dataset, ktorý by bol využiteľný na všetky účely skúmania v oblasti digitálnej forenznej analýzy. Pri používaní, vytváraní a zdieľaní datasetov existuje niekoľko výziev, s ktorými sa výskumníci stretávajú. Dôležité je poznamenať, že vo všeobecnosti takéto datasety v tejto oblasti chýbajú, alebo chýba dokumentácia a formálny popis jeho konštrukcie. Na účely nášho výskumu potrebujeme pracovať s datasetom, ktorý popisuje reálne scenáre, ktoré môžu nastať v rámci skutočných bezpečnostných incidentoch.

2.1.1 Aktuálny stav riešenej problematiky

Výzvou v oblasti forenznej analýzy je nedostatok kvalitných datasetov, ktoré by spĺňali viacero kritérií, ako je kvalitná prípadová štúdia a dostatok záznamov (artefaktov) [16, 17]. Niekoľko článkov a štúdií sa zaoberalo touto problematikou.

V dvoch z článkov [15, 17] sa autori zaoberali otázkou existencie dostatočne kvalitných datasetov pre digitálnu forenznú analýzu a popísali, ako by takýto dataset mal byť vhodne zostavený. Zároveň zdôraznili nedostatok zdieľaných datasetov v tejto oblasti.

Ďalší aspekt spojený s analýzou digitálnych stôp a artefaktov je ontologický zápis jednotlivých artefaktov, čo môže pomôcť analyzovať tieto artefakty a nájsť medzi nimi vzájomné súvislosti. V niektorých vedeckých prácach sa uvádza štandardný digitálny forenzný formát s názvom štandard CASE (Cyber-investigation Analysis Standard Expression), ktorý slúži na štandardizáciu digitálnej forenznej analýzy [18, 19, 20].

V inom článku [21] autori vo svojich zisteniach poukazujú na to, že pri vytváraní datasetu netreba zabúdať na to, aby sa datasety nielen že vytvorili korektne, ale taktiež odrážali také scenáre, ktoré sa môžu vyskytnúť v reálnych situáciách. Tiež vyzdvihujú, že vytváranie a zdieľanie datasetov môže byť výzvou z dôvodu podliehania zákonným obmedzeniam alebo správe a ochrane údajov. Jednak je potrebné buď vhodne reprezentovať údaje v datasete, alebo môže dôjsť k zdieľaniu datasetu, pričom autor si nemusí byť vedomý toho, že má isté povinnosti v súvislosti s obsahom datasetu.

Grajeda a spol. [16] sa venujú dostupnosti datasetov v tejto oblasti, ale tiež sa zameriavajú na dôležitosť zdieľania datasetov z dôvodu replikácie výsledkov a tiež na zlepšenie výskumu. Datasetov vo všeobecnosti existuje viacero, avšak mnoho z nich sa nezameriava na digitálnu forenznú analýzu ako takú, ale na jednu z oblastí – emailová komunikácia, malvér, sieťová prevádzka a podobne. Ukázalo sa, že mnoho výskumníkov sa rozhodlo datasety nezdieľať a to hneď z niekoľkých dôvodov. Prvým je nedostatočná možnosť zdieľania datasetu (napríklad z dôvodu dostupných, resp. nedostupných zdrojov alebo z dôvodu chýbajúcej stabilnej platformy, cez ktorú by ho bolo možné zdieľať). Ďalšími dôvodmi sú práve obava z porušenia rôznych práv (napr. autorských práv) a podobne alebo si v čase písania článku neuvedomili potrebu zdieľania datasetu.

Autori Breitinger a Jotterand [22] vo svojom výskume popisujú novú taxonómiu, ktorú by mali výskumníci pri tvorbe datasetu používať na štruktúrovanie údajov a tiež sa venujú právnym aspektom zdieľania údajov. Skonštatovali, že vytváranie a zdieľanie datasetov je nevyhnutné pre pokrok a umožnenie porovnávania výsledkov. Zdôraznili tiež, že je potrebné byť opatrný z hľadiska osobitných zákonov napríklad autorské právo alebo licencovanie. Keďže článok je z roku 2023, je zjavné, že datasety využívané vo výskumoch sa stále nie vždy zdieľajú, alebo môžu porušovať konkrétnu právnu úpravu.

V dnešnej dobe existujú isté rámce, ktoré umožňujú vytváranie a generovanie datasetov v oblasti digitálnej forenznej analýzy. Jedným z prvých rámcov je Forensig2 [23, 24], ktorý slúži na vytváranie obrazov diskov. Používateľ naprogramuje v skripte správanie používateľa v systéme a vygeneruje obraz disku, ktorý je možné následne analyzovať. Ďalším rámcom je ForGe s grafickým rozhraním [25], ktorý je schopný vytvárať obraz disku so súborovým systémom NTFS. Scanlon a spol. navrhli EviPlant [26], čo je systém určený na efektívne vytváranie, manipuláciu, ukladanie a distribúciu výziev pre vzdelávanie a školenia v oblasti digitálnej forenznej analýzy. Ďalším

navrhnutým rámcom je TraceGen, [27], čo je automatizovaný systém zameraný na emuláciu akcií používateľa s cieľom vytvoriť realistické a komplexné artefakty kontrolovateľným a reprodukovateľným spôsobom. Posledným, nami spomenutým, rámcom je ForTrace [28], ktorý je schopný simultánneho generovania trvalých, volatilných a sieťových stôp.

2.1.2 Prístup k riešeniu výskumného cieľa

Cieľom je vytvorenie vhodného datasetu pre porovnávanie metód pri analýze digitálnych stôp, ktorý by bolo možné používať na skúmanie rôznych problémov. Na základe vyššie spomenutých výskumných prác môžeme skonštatovať, že vo všeobecnosti existujú rôzne datasety, s ktorými výskumníci pracujú, avšak ich buď nezdieľajú, alebo nie všetky sú využiteľné v rámci nášho výskumu. Takisto existujú rámce, pomocou ktorých je možné datasety generovať, neexistuje však istota, ktorá by zaručovala, že dataset je korektne vytvorený a spĺňa všetky podmienky potrebné pre náš výskum.

V našom výskume pracujeme s časťou datasetu, ktorý sme vykonštruovali z dát dostupnými z portálu DFIR Madness. Použitý bol prípad s názvom „Case 001 – The Stolen Szechuan Sauce“ (bližšie ho popisujeme v kapitole 3). Ďalším plánovaným krokom je popísať a publikovať časť datasetu, ktorý by mohol prispieť svojou existenciou výskumu v oblasti digitálnej forenznej analýzy vzhľadom k nedostatku zdieľaných datasetov.

Ako sme vyššie spomenuli, vo všeobecnosti chýbajú datasety, ktoré by boli využiteľné v rámci výskumu. Je totiž bežnou praxou, že zverejnené datasety len málokedy odzrkadľujú realitu a reálne scenáre, ktoré môžu nastať pri skutočných incidentoch. Datasety, ktoré sa publikujú z takzvaných CTF („Capture the flag“) súťaží by mohli tento problém odstrániť, avšak aj to je stále víziou budúcnosti. V CTF súťažiach je väčšinou odsimulovaný nejaký konkrétny útok, ktorí súťažiaci analyzujú. Zverejnenie práve takýchto datasetov by mohlo pomôcť oblasti digitálnej forenznej analýzy.

Jednou z najdôležitejších otázok je teda dôvod, prečo chceme použiť údaje, ktoré sú z reálneho (alebo aspoň simulovaného) incidentu. Pracovali sme aj s myšlienkou generovania náhodných dát, avšak po otestovaní sme došli k záveru, že ak na vstupe dáme náhodné údaje, na výstupe nemôžeme získať nič iné, len náhodné výsledky.

Okrem publikovania datasetu, s ktorým pracujeme, uvažujeme nad vytvorením syntetického datasetu, ktorý by sme mohli považovať za zovšeobecnenie už existujúcich datasetov. Hlavnou myšlienkou je zozbierať existujúce datasety z už ukončených CTF súťaží. Následne je dôležité označiť, ktoré záznamy sú benígne a ktoré škodlivé, teda anomálne. V ďalšom kroku sme schopní synteticky generovať benígne a anomálne záznamy. Na zachovanie kódovania uvažujeme použiť asociačné pravidlá, ktoré nám poskytuje metóda nazývaná formálna konceptová analýza (FCA), ktorú sme použili v našich predošlých výstupoch [29, 30]. Asociačné pravidlá nám budú slúžiť ako vstup a budú popisovať spôsob, ako vytvárať a generovať údaje. Atribúty pre každý záznam musia byť v binárnom tvare, pretože ak pri syntetickom generovaní záznamov budeme mať niektoré atribúty nastavené na 1 (pre konkrétny záznam v datasete), tak pomocou asociačných pravidiel budeme vedieť dogenerovať zvyšné atribúty tak, aby daný záznam bol buď benígny, alebo anomálny. Zaujímavou otázkou pri použití asociačných pravidiel na generovanie záznamov je, či sú schopné zachovať časový sled potenciálneho útoku.

2.2 Identifikácia relevantných digitálnych stôp

Druhý cieľ súvisí s identifikáciou relevantných digitálnych stôp v systéme a nájdenie vhodného spôsobu ako vyhodnotiť výsledky vybraných použitých metód. Vo všeobecnosti môžeme tento výskumný cieľ rozdeliť na dva podciele a to:

1. Nájdenie vhodných metód na identifikáciu relevantných digitálnych stôp.
2. Nájdenie vhodného spôsobu, ako vyhodnotiť nadobudnuté výsledky vybraných metód strojového učenia.

Táto problematika je esenciálnou v oblasti digitálnej forenznej analýzy. Na to, aby bol bezpečnostný incident riešený v čo najkratšom čase, je potrebné, aby boli čím skôr identifikované relevantné digitálne stopy. Je to základná činnosť pri riešení bezpečnostného incidentu, bez ktorej analytik nemôže pokračovať vo fázach digitálnej forenznej analýzy. Mnohí výskumníci a forenzní vyšetrovatelia sú si vedomí výhod čiastočnej automatizácie rôznych fáz digitálnej forenznej analýzy, preto sa snažia nájsť rôzne spôsoby, ako by automatizáciu implementovali do týchto procesov.

V nasledujúcich podkapitolách sa zaoberáme relevantnými článkami týkajúcimi sa výskumu identifikácie anomálií v digitálnych stopách alebo forenzných artefaktoch.

Tieto súvisiace práce sú rozdelené do troch kategórií: články zamerané na záznamy (logy), súborové systémy a registre. Takisto sa venujeme prácam, ktoré sa venujú spôsobu vyhodnotenia metód použitých na hľadanie anomálií, teda v našom prípade na identifikáciu relevantných digitálnych stôp.

2.2.1 Aktuálny stav riešenej problematiky

Automatizácia je v súvislosti s digitálnou forenznou analýzou kľúčová, pretože môže urýchliť rôzne fázy vyšetrovacieho procesu. V článku autorov Du a Scanlona [10] bola navrhnutá metodológia na automatické uprednostňovanie podozrivých súborových artefaktov, aby sa znížila potreba manuálnej analýzy. Autori predstavili aj sadu nástrojov na extrahovanie údajov z obrazov diskov. Skopik a spol. [31] študovali metódy analýzy záznamov (logov) pomocou algoritmov strojového učenia na detekciu anomálií. Beebe a spol. [32] zistili, že škodlivú aktivitu insiderov v rámci zariadení možno identifikovať analýzou údajov a forenzných artefaktov súvisiacich so súbormi. Ich výskum poskytol výpočtový prístup na nájdenie digitálnych forenzných stôp, pričom manuálna analýza potvrdila väčšinu zistených anomálií.

Prvá skupina článkov, ktorým sme sa venovali, sa zameriava na **detekciu anomálií v záznamoch operačného systému**. Xu a spol. v ich článku [33] analyzovali konzolové záznamy a aplikovali metódy na detekciu anomálií, pričom PCA už v roku 2009 priniesla dobré výsledky. Studiawan a spol. v štúdii [34] zdôraznili význam záznamov udalostí (event logov) ako cenného zdroja digitálnych stôp pre forenzné vyšetrovanie, pretože zaznamenávajú základné systémové aktivity. Vo svojom ďalšom príspevku Studiawan a spol. v článku [35] navrhli metódu klastrovania záznamov riadenia prístupu pomocou algoritmu MajorClust a identifikácie anomálií. V roku 2020 autori v článku [36] tiež vyhodnotili výkon siedmich detektorov anomálií v systémových záznamoch. Vo svojom výskume z roku 2021 sa Studiawan a spol. [1] zamerali na detekciu anomálií pomocou hlbokých autoenkóderov. Ďalší výskumníci tiež použili hlboké autoenkódery na detekciu anomálií, vrátane Liu a spol. [37] na odhaľovanie vnútorných hrozieb a Yuan a spol. [38] na detekciu anomálnych používateľov. Hu a spol. [39] taktiež zisťovali anomálne správanie používateľa, ale pomocou záznamov z viacerých zdrojov. V inom výskume He a spol. [40] preskúmali a diskutovali o troch metódach detekcie anomálií s učiteľom a bez učiteľa. Hirakawa a spol. [41] navrhli

metódu detekcie anomálií založenú na riedkych atribútoch a vnútornom stave modelu, o ktorej tvrdia, že funguje dobre, aj napriek malému množstvu trénovacích údajov.

Druhá skupina výskumných prác sa zameriava na **detekciu anomálií v súborových systémoch**, keďže väčšina digitálnych stôp je uložená v súborovom systéme počítača. V roku 2005 Carrier a spol. [42] preskúmali možnosť automatizácie vyhľadávania súborov a adresárov vytvorených počas incidentu. Pirker a spol. [43] sa zamerali na odhaľovanie nových útokov porovnaním správania sa v systéme. Použili kombináciu zberu udalostí a analýzy údajov na „odtlačky“ procesov a ich správanie pri prístupe k súborom. Du a spol. [44] zaviedli metódu hodnotenia relevantnosti artefaktov súborov, ktorá sa opiera o centralizovaný model DFaaS (Digital Forensics as a Service). Využitím relevantných súborov, s ktorými sa pracovalo pri predchádzajúcich vyšetrovaniach, ich metóda dokáže klasifikovať novoobjavené súbory.

Ďalšia skupina výskumných prác sa zameriava na **detekciu anomálií v registroch**. Stolfo a spol. sa v článku [45] pokúšali zistiť nezvyčajnú aktivitu v registroch operačného systému Windows pomocou metódy Support Vector Machines. V inom článku autori [46] navrhli prístup nazývaný RAMD (registry-based anomaly malware detection), ktorý používa súborový klasifikátor na detekciu škodlivého softvéru, ktorý zneužíva kľúče registra. Chouhan a spol. v ich článku [47] odhaľovali anomálne správanie v rámci procesov skúmaním anomálií v registroch, systémových knižniciach a prístupov k súborom procesu.

Je dôležité sa zaoberať aj spôsobom, ako vyhodnotiť dosiahnuté výsledky. Vo všeobecnosti je problematika vyhodnocovania metód na hľadanie anomálií netriviálny problém. Vzhľadom k tomu, že využívame metódy na hľadanie anomálií práve pri riešení bezpečnostného incidentu (a teda bez učiteľa), v datasete sa nachádzajú neoznačkované údaje. Preto je potrebné uvažovať nad tým, či je vhodné brať do úvahy všetky hodnoty pre falošne pozitívne (FP), falošne negatívne (FN), pravdivo pozitívne (TP) a pravdivo negatívne (TN) triedy.

V článku [48] sa autori zameriavajú na vyhodnocovanie metód na detekciu anomálií. Porovnávajú výsledky pre 12 metód nad 5 rôznymi datasetmi. Popisujú, že vzhľadom k tomu, že väčšina datasetov je pri hľadaní anomálií nevyvážená (je v ňom viac inlierov ako outlierov), Precision, Recall a F1 skóre sú metriky, ktoré sa bežne používajú na meranie výkonnosti modelov. Na evaluáciu navrhujú použiť práve tieto tri

metriky a tiež „Area under the precision-recall curve“ (AUPR) a „Area under the ROC curve“ (AUC), keďže sú citlivejšie na predikciu pozitívnej triedy, čo je vhodnejšie na použitie pri detekcii anomálií.

V ďalšom článku [49] autori spomínajú, že F1 skóre a „Average Precision“ (AVPR) sú citlivé na nastavenie hodnôt pre parameter kontaminácia a teda by nemali byť použité pri vyhodnocovaní metód na detekciu anomálií. Namiesto toho odporúčajú použiť iné metriky ako napríklad AUC.

Goldstein a Uchida [50] vo svojom výskume tvrdia, že porovnanie výsledkov metód na detekciu anomálií bez učiteľa môže byť nejednoznačné. Odporúčajú využívať metriky ako AUC alebo integrál „Receiver operator characteristic“ (ROC). AUC je v tomto prípade pravdepodobnosť, že algoritmus detekcie anomálií pridelí náhodne vybranému normálnemu prípadu nižšie skóre ako náhodne zvolenej anomálnej inštancii. Na druhej strane neodporúčajú AUC použiť pri problémoch s nevyváženými triedami.

Zoppi a spol. [51] navrhli nástroj RELOAD (Rapid EvaLuation Of Anomaly Detection algorithms), ktorý je schopný spustiť niekoľko metód na hľadanie anomálií a tiež poskytuje výsledky na základe rozsiahleho súboru metrík spolu s vizualizačnými prvkami. Autori sledujú hodnoty pre TP, TN, FP, FN, Precision, Recall, False Positive Rate, Accuracy, F1 skóre, Mathewov koeficient (MCC) a tiež AUC.

V poslednom článku [52] autori popisujú vyhodnocovanie rôznych metód na detekciu anomálií avšak s učiteľom. Pri klasifikácii však tiež využívajú na vyhodnotenie metriky Accuracy, Precision, Recall, False Positive Rate, F1 skóre, ROC a tiež AUC pre každú z 12 nimi vybraných metód.

2.2.2 Prístup k riešeniu výskumného cieľa

Druhý cieľ je identifikácia relevantných digitálnych stôp. Jedná sa o veľmi dôležitú časť nášho výskumu z dôvodu vo všeobecnosti chýbajúcej rýchlej reakcie na bezpečnostné incidenty.

Mnoho predchádzajúcich štúdií zdôrazňovalo používanie metód s učiteľom, pretože sa spoliehajú na označkované datasety a predpokladajú určitú úroveň predchádzajúcich informácií o údajoch. Na rozdiel od toho sa náš prístup zameriava na

neohodnotené datasety, pretože počas vyšetrovania bezpečnostného incidentu nemusí byť identifikovaný počiatočný bod vyšetrovania a relevantné digitálne stopy okamžite zrejmé.

V našom výskume sa zaoberáme metódami bez učiteľa na identifikáciu anomálií a relevantných digitálnych stôp a na hľadanie vzťahov. V publikovaných článkoch [29, 30, 53] sa konkrétne venujeme metóde LOF na hľadanie outlierov a formálnej konceptovej analýze. Formálnu konceptovú analýzu bližšie popisujeme v kapitole 2.3. V ďalšom výskume sme sa zamerali aj na ďalšie metódy na hľadanie outlierov, pričom parciálne výsledky uvádzame a porovnávame v kapitole 4.

Taktiež už z podobných výskumných prác je vidieť, že neexistuje jednoznačné odporúčanie na vyhodnocovanie metód na detekciu anomálií. Niektorí tvrdia, že je vhodné použiť F1 skóre, Precision a Recall, iní zas, že je vhodnejšie použiť napríklad metriku AUC. Aktuálne v našom výskume využívame na vyhodnotenie práve F1 skóre, Precision a Recall, avšak pri plnení cieľa dizertačnej práce sa chceme zamerať aj na iné metriky, ktoré odporúčajú ďalší výskumníci ako napríklad AUC alebo AUPR.

Aktuálne pristupujeme k riešeniu tohto výskumného cieľa tak, že využívame na vyhodnocovanie metód strojového učenia maticu zmeny a teda sledujeme hodnoty TP, TN, FP, FN, Precision, Recall a F1 skóre. Accuracy sme z dôvodu nevyváženého datasetu úplne vynechali. Parciálne výsledky pre rôzne metódy popisujeme v kapitole 4.

Ďalším krokom pri spĺňaní tohto cieľa je pozrieť sa na výsledky pre metriky AUC, ROC, AUPR, MCC a sledovať ich správanie pri nastavovaní rôznych parametrov vybraných metód na hľadanie anomálií. Takisto chceme zistiť, ktoré metriky na vyhodnocovanie metód je naozaj vhodné použiť nad údajmi, ktoré využívame v našom výskume, keďže pri našom probléme pracujeme s datasetom, ktorý je nevyvážený (nepomer medzi pozitívnou a negatívnou triedou).

2.3 Hľadanie vzťahov medzi digitálnymi stopami

Posledným cieľom práce je hľadanie vzťahov medzi digitálnymi stopami a ich atribútmi pomocou metód strojového učenia. Keďže analytik vo všeobecnosti manuálne hľadá vzťahy medzi digitálnymi stopami, aby mohol vyvodzovať závery, chceme týmto výskumom urýchliť práve tento proces. Vo všeobecnosti je nájdenie vzťahov medzi digitálnymi stopami netriviálny problém.

2.3.1 Aktuálny stav riešenej problematiky

Pri riešení problému hľadania vzťahov či už medzi digitálnymi stopami alebo vo všeobecnosti v oblasti digitálnej forenznej analýzy sa vyskytuje niekoľko rôznych metód, ktoré sú využívané. Medzi ne patria napríklad zhlukovanie (clustering), formálna konceptová analýza, alebo riešenia založené na grafovej terminológii.

Prvá oblasť, ktorou sme sa zaoberali sú grafy a ich využitie v oblasti digitálnej forenznej analýzy. Vo všeobecnosti je grafy možné v rámci digitálnej forenznej analýzy použiť na množstvo interpretácií. V prvom článku [54] sa autori venovali porovnaniu grafovej teórie a váhových metód pri obnove súborov z povrchu disku na základe jeho štruktúry (file carving), pričom sa využívajú rôznorodé metódy a prístupy. V tejto štúdii boli využité metadáta a štatistické postupy. Okrem toho sa autori venovali aj technikám fragmentácie súborov a algoritmom obnovy, s cieľom analyzovať efektívnosť procesu obnovy pre rozličné formáty súborov. Zistenia vyplývajúce z analýzy poukazujú na to, že predchádzajúce výskumy inklinujú k teoretickým prístupom z oblasti grafov, konkrétne váhovým metódam, ako prostriedku systematického manévrovania veľkým objemom údajov spojených so súborovou fragmentáciou. V ďalšom článku [55] autori predstavujú aplikáciu Temporal Analysis Integration Management Application (TAIMA) s cieľom vylepšiť proces digitálnej forenznej analýzy využitím metód vizualizácie informácií. TAIMA predstavuje prototypovú aplikáciu, ktorá poskytuje časovo orientovaný grafický rámec pre rekonštrukciu udalostí prostredníctvom abstrakčných a vizuálnych postupov. Pomáha pri identifikácii kľúčových systémových udalostí v rámci procesu digitálnej forenznej analýzy. Na hľadanie vzťahov medzi digitálnymi stopami v ďalšom článku [56] autori využívajú grafové neurónové siete, aby tým pomohli objaviť či už vzťahy alebo vzory v digitálnych stopách.

Druhou skupinou článkov sú výskumné práce zaoberajúce sa zhlukovaním v oblasti digitálnej forenznej analýzy. V jednom z článkov [57] autori navrhli rámec, ktorý využíva kernelové k-means v kombinácii s metódou podporných vektorov na analýzu dokumentov a obrázkov v rámci forenzného vyšetrovania. Autori ďalšieho článku [58] navrhujú použitie zhlukovacích a klasifikačných algoritmov na dosiahnutie inteligencie v procese digitálneho forenzného vyšetrovania. Tiež prichádzajú s komplexným blokovým diagramom ako všeobecným rámcom pre inteligentné forenzné vyšetrovanie. Posledný článok [59] sa zameriava na problém triáže z dôvodu lokalizácie digitálnych stôp počas automatického digitálneho forenzného vyšetrovania. V rámci

výskumu autori využívajú zhlukovací algoritmus založený na stratégii transformácie rozsahu, ktorý je schopný identifikovať podobné minulé prípady obsahujúce stopy, ktoré sú znovu použiteľné.

Tretím prístupom je použitie formálnej konceptovej analýzy. Okrem nášho výskumu [29, 30], táto metóda je na hľadanie vzťahov medzi digitálnymi stopami vo všeobecnosti len veľmi málo používaná. Avšak zo všeobecného hľadiska sa formálna konceptová analýza používala v niekoľkých článkoch. V prvom z článkov [60] sa autori venovali analýze digitálnych stôp pri útokoch pomocou sociálneho inžinierstva s použitím formálnej konceptovej analýzy. Iný výskum [61] sa zameriava na vytvorenie zmysluplnej reprezentácie techník a postupov hrozieb s cieľom poskytnúť odborníkom prostriedky na včasnejšiu a strategickú detekciu kybernetických hrozieb. Autori to dosiahli aplikáciou formálnej konceptovej analýzy na existujúci taxonomický rámec MITRE ATT&CK s cieľom získať hlbšie konceptualizácie a vzájomné vzťahy. V poslednom z článkov [62] autori navrhli vyšetrovací proces prostredníctvom vizualizácie a analýzy údajov mobilných telefónov pomocou formálnej konceptovej analýzy. Autori vizualizujú konceptový zväz, ktorý môže byť vnímaný ako súbor spoločných a odlišných atribútov údajov.

Hľadaniu vzťahov v oblasti informačnej a kybernetickej bezpečnosti sme sa venovali aj my v predošlom výskume. Určitou inšpiráciou práve pre hľadanie vzťahov v oblasti digitálnej forenznej oblasti boli aj dva naše výstupy [63, 64]. V prvom [63] sme sa venovali klasifikácii škodlivých emailových správ do troch kategórií – spam, scam a phishing. Extrahovali sme z emailov 11 rôznych atribútov a implementovali sme štyri metódy strojového učenia s učiteľom. V druhom výstupe [64] sme použitý dataset obohatili o ďalšie emaily, kategórie rozšírili o novú skupinu – emailové správy obsahujúce malvér a takisto extrahovali viac relevantných atribútov pre škodlivé emailové správy.

2.3.2 Prístup k riešeniu výskumného cieľa

Vo všeobecnosti existuje len málo výskumných prác, ktoré by sa primárne zaoberali len hľadaním vzťahov medzi digitálnymi stopami a ich atribútmi. Celkovo výskumné práce riešia tento problém ako vedľajší, skôr sa zameriavajú na iné problematické oblasti v rámci digitálnej forenznej analýzy.

V našom výskume sa aktuálne zameriavame na použitie formálnej konceptovej analýzy na získanie prehľadu o tom, aké vzťahy existujú medzi digitálnymi stopami alebo ich atribútmi. Formálna konceptová analýza (FCA) je metóda používaná najmä na analýzu údajov a teda na odvodenie implicitných vzťahov medzi objektmi opísanými prostredníctvom súboru atribútov. Pracovali sme so záznamami súvisiacimi so súborovým systémom NTFS a takisto s evtx súbormi. Sledovali sme napríklad koreláciu medzi atribútmi súvisiacimi s časovými pečiatkami (M, A, C, B) a takisto s ďalšími atribútmi, ktoré by mohli indikovať podozrivé správanie v systéme. Nad skúmaným datasetom sme vytvorili asociačné pravidlá, ktoré nám popisujú vzájomné vzťahy medzi atribútmi záznamov. Vzhľadom k tomu, že pracujeme s binárnymi atribútmi (čo je bližšie popísané v kapitole 3), bolo jednoduché tieto asociačné pravidlá vygenerovať. Podrobnejšie naše dosiahnuté výsledky popisujeme v publikovaných článkoch [29, 30].

Ďalším možným smerovaním je práve použitie zhlukovacích algoritmov, ktoré by mohli byť schopné rozdeliť záznamy do rôznych skupín na základe hodnôt atribútov. Pri vytváraní zhlukov sa v každej skupine nachádzajú záznamy s podobnými atribútmi, preto by bolo zaujímavé skúmať, akým spôsobom sú tieto záznamy rozdelené. Vzhľadom k tomu, že zhlukovanie je úplne iná metóda strojového učenia ako hľadanie anomálií, je tiež potrebné zamýšľať sa nad spôsobom vyhodnocovania dosiahnutých výsledkov.

3 Identifikácia relevantných digitálnych stôp pomocou metód na hľadanie outlierov

Náš výskum je obmedzený na operačný systém Windows a jeho predvolený súborový systém NTFS (New Technology File System). Na základe toho navrhujeme spôsob, ktorým identifikujeme digitálne stopy relevantné pre daný prípad. Týmto riešením je možné uľahčiť prácu forenzným analytikom najmä v počiatočných fázach ich analytickej činnosti (po fáze extrakcie forenzných artefaktov).

Dôležitým aspektom digitálnej forenznej analýzy je identifikácia relevantných digitálnych stôp pre analyzovaný prípad. Zamerali sme sa na metódy detekcie outlierov bez učiteľa z dôvodu práce s neohodnoteným datasetom. Outliera možno definovať ako „pozorovanie v datasete, ktoré sa zdá byť nekonzistentné so zvyškom tohto datasetu“ [65]. Z forenzného hľadiska bolo potrebné hľadať anomálne súbory a názvy súborov, preto je náš výskum rozdelený na dve časti. Prvá časť je analýza inodov súborov a druhá časť je analýza názvov súborov, pričom za príslušné digitálne stopy považujeme práve relevantné inody a názvy súborov.

Nájdenie vhodného datasetu je jednou z významných úloh v oblasti výskumu digitálnej forenznej analýzy [16]. Podľa autorov článku [21] rozlišujeme tri typy datasetov v oblasti digitálnej forenznej analýzy: datasety na hodnotenie nástrojov/procesov, datasety akcií a datasety založené na scenároch. Mnohé dostupné datasety predstavujú prvé dve kategórie.

Datasety založené na scenároch sa často generujú na simuláciu konkrétneho útoku, kybernetického incidentu alebo potenciálneho vyšetrovacieho príbehu. Pre náš výskum je vhodná práve táto kategória datasetov. V článku [21] autori popisujú požiadavky na tieto datasety. Za zmienku stojí najmä zoznam všetkých aktivít vykonaných na zariadení/v systéme a všetky súbory s dôkaznou hodnotou. Keďže cieľom nášho výskumu je hľadať relevantné digitálne stopy, takéto zoznamy sú nevyhnutné. Z týchto dôvodov pre tento výskum využívame modelový prípad z portálu DFIR Madness s názvom Case001 – The Stolen Szechuan Sauce [66].

V nasledujúcich podkapitolách popisujeme ukážku prístupu riešenia. Pre pochopenie kontextu popisujeme prípad, dataset a predspracovanie datasetu, čomu sme

sa venovali už v predošlých výskumoch [29, 30, 53]. Tiež popisujeme použité metódy a metriky používané pri evaluácii.

3.1 Popis modelového prípadu

Prípad je o tajnom recepte sečuánskej omáčky od firmy CITADEL, ktorý sa našiel na darkwebe, pričom hlavnou úlohou je zistiť ako sa tam recept dostal. Spoločnosť požiadala o forenznú analýzu svojho doménového kontroléra a hostiteľa siete. Doplnkové úlohy zahŕňajú otázky, ako napríklad, či došlo k úniku údajov, inicializačný vektor postupu útočníka, či bol použitý malvér a podobne. Dôležitou otázkou je, či vôbec útočník naozaj ukradol recept na sečuánsku omáčku a či sa dostal k ďalším citlivým údajom. Portál DFIR Madness poskytuje nasledujúce digitálne stopy na účely forenznej analýzy:

- Obraz disku DC01 (E01)
- Pamäť a pagefile z DC01
- Autoruns z DC01
- Protected Files z DC01
- PCAP
- Obraz disku Desktop-u (E01)
- Pamäť a pagefile z Desktop-u
- Autoruns z Desktop-u
- Protected Files z Desktop-u;

V rámci nášho výskumu sme sa zamerali len na obraz disku pre DC01 (E01). Pre popis prípadu sme identifikovali operačný systém analyzovaných zariadení. Pre server to bol operačný systém Windows a na klientskej stanici sa nachádzal Windows 10. Dôležitou informáciou je aj lokálny čas servera, ktorý bol nastavený na Mountain Standard Time (UTC -6). Aby sme prípadu poskytli viac kontextu, analyzovali sme aj sieťovú prevádzku, konkrétne súbor pcap - Case001.pcap. Analýzou tohto súboru sme zistili, že sa jednalo o sken IP adries spoločnosti CITADEL dňa 19.09.2020. Po skenovaní

došlo k útoku hrubou silou na Remote Desktop Service (RDP) z IP adresy 194.61.24.102, po ktorom sa útočník úspešne prihlásil pod účtom Citadel/Administrator. Tieto udalosti sa zaznamenávajú do záznamov servera - Security EVTX. Zistili sme, že pomocou rovnakých prihlasovacích údajov útočník vytvoril ďalšiu reláciu RDP z doménového kontroléra do klientskeho zariadenia. Cez Internet Explorer bol neskôr stiahnutý a nainštalovaný malvér s názvom coreupdater.exe na server a klientske zariadenie. Identifikovali sme ho ako Metasploit schopný napríklad migrácie procesov, krádeže prihlasovacích údajov, zaznamenávania kláves a obrazovky. Záznam o tejto aktivite sa nachádza aj v sieťovej prevádzke a kľúčoch registra servera aj klienta. Stopy o vytvorení, úprave a vymazaní pôvodných súborov sa nachádzajú v záznamoch MFT, záznamoch vo všeobecnosti (logoch) a súboroch .lnk. Výsledkom analýzy .lnk súborov je prístup k súboru SzechuanSauce.lnk dňa 19.9.2020 o 03:32:21 (UTC -6). Analýza záznamov potvrdila, že došlo aj k úniku údajov. Exfiltrácia údajov bola zachytená v súbore Secret.zip, ktorý obsahuje súbor tajného receptu „Szechuan sauce“.

3.2 Dataset a predspracovanie

Táto kapitola sa zaoberá popisom predspracovania údajov a vytvoreného datasetu. Ako bolo uvedené vyššie, v našom výskume sme sa obmedzili na obraz disku predmetného servera. Fáza predspracovania údajov pozostáva z vytvorenia časovej osi, extrakcie údajov zo súborového systému a jeho úpravy a obohatenia. Na vytvorenie časovej osi sme použili nástroj Log2timeline (plaso). Je to nástroj, ktorý umožňuje vytváranie časových línií a má veľa doplnkov, syntaktických analyzátorov a analyzátorov, ktoré vykonávajú predbežné načítanie digitálnych udalostí obsiahnutých v zariadení.

Keďže sme analyzovali obraz disku zariadenia so systémom Windows, vybrali sme analyzátor win7_slow, ktorý obsahuje tri ďalšie analyzátory, konkrétne win_gen, webhist a win7. Výstupom nástroja Log2timeline je časová os v nečitateľnej podobe, preto sme na ňu aplikovali ďalší z nástrojov plaso, a to nástroj psort.py, ktorý previedol výstup z predchádzajúceho kroku do čitateľnej podoby. Ako výstupný formát súboru sme zvolili formát l2tcsv, čo je jednoduchý súbor CSV so 17 poliami, ktoré tvoria jeho hlavičku. Ostatné riadky v súbore predstavujú jednotlivé záznamy, každý so svojou časovou pečiatkou.

Výslednú časovú os sme ďalej upravili pomocou programovacieho jazyka Python verzie 3 a knižnice pandas. Počiatočný počet záznamov pred rozdelením bol 1 256 180. Časová os obsahovala záznamy z jedenástich rôznych zdrojov údajov. Zdroje údajov a ich zodpovedajúci počet záznamov sú uvedené v Tab. 1.

Tab. 1 Rozdelenie záznamov digitálnych stôp podľa zdroja

Zdroj	**Počet záznamov**
AMCACHE	136
AMCACHEPROGRAM	3
EVT	86 180
FILE	843 863
LNK	45
LOG	194
OLECF	253
PE	18 115
RECBIN	1
REG	307 315
WEBHIST	75
Všetky	1 256 180

Najvýraznejšie zastúpenie mali údaje zo zdrojov FILE, EVT a REG, ktoré tvorili 87% všetkých záznamov. Ďalšia úprava údajov zahŕňala odstránenie záznamov obsahujúcich nulovú časovú pečiatku a rozdelenie údajov do 11 samostatných údajových rámcov podľa zdroja. V tomto príspevku sme použili iba údaje, ktorých zdrojom bol súborový systém a počet záznamov bol 843 863.

V ďalšom kroku boli z primárnych polí každého záznamu extrahované rôzne kategorické a binárne atribúty. Kategorické atribúty boli tiež prevedené na binárne. Atribúty boli potom rozdelené do siedmich kategórií:

- Atribúty súvisiace s časovými pečiatkami: timestamp = 'M', 'A', 'C', 'B';
- Atribúty súvisiace so zdrojom: source_type = 'file_stat', 'NTFS_file_stat'; 'file_entry_shell_item', 'NTFS_USN_change';
- Atribúty súvisiace s tým, či sa jedná o súbor, adresár, alebo odkaz na súbor: file_type = 'filef', 'directory', 'link';
- Atribúty súvisiace s typom adresára: dir_type = 'dir_appdata', 'dir_win', 'dir_user', 'dir_other';
- Atribúty súvisiace s príponou súboru: file_type2 = 'file_executable', 'file_graphic', 'file_documents','file_ps', 'file_other';
- Atribúty súvisiace s formátom súboru: file_format = 'mft', 'lnk_shell_items', 'olecf_olecf_automatic_destinations/lnk/shell_items', 'winreg_bagmru/shell_items', 'usnjrnl';
- Atribúty súvisiace s veľkosťou súboru: file_size = 'size_none', 'size_Q1', 'size_Q2', 'size_Q3', 'size_Q4'.

Ďalším krokom bolo zúženie datasetu na čas bezpečnostného incidentu, ktorý bol stanovený na čas od 22:24:50 18. septembra 2020 (UTC) do 4:52:45 19. septembra 2020 (UTC). Tento krok znížil počet záznamov na 9860. Vykonali sme manuálnu digitálnu forenznú analýzu, aby sme overili vybrané metódy a identifikovali relevantné digitálne stopy, a teda relevantné názvy a inody súborov. Identifikovali sme 17 názvov súborov a 15 inodov súborov, ktoré boli relevantné pre prípad:

- **Inody súborov**: 84630, 84880, 84987, 86966, 86967, 86968, 86970, 86971, 86975, 87059, 87060, 87064, 87111, 87112 a 87137;
- **Názvy súborov**: "coreupdater.exe", "FILESH~1", "Secret", "BETH_S~1.TXT", "Beth_Secret.lnk", "SECRET~1.TXT", "SECRET_beth.lnk", "Szechuan", "SZECHU~1.TXT", "Secret.lnk", "NoJerry.lnk", "NoJerry.txt",

"f01b4d95cf55d32a.automaticDestinations-ms", "SECRET_beth.txt", "Beth_Secret.txt", "Secret.zip" a "coreupdater.exe.2424urv.partial".

Analyzovali sme atribút inode a našli sme tie najbežnejšie v datasete (0 a 84656), ktoré sme vynechali. Tieto inody súvisia s metasúbormi $MFT a $UsnJrnl. Po vykonaní tohto kroku sme získali 665 záznamov relevantných pre bezpečnostný incident v rámci výskumu, ktorý sa zameriaval na inode. Podobnú úlohu sme vykonali aj s názvami súborov, pričom sme vynechali „NTFS:$MFT“ a „setupapi.dev.log“, výsledkom čoho bolo 9279 záznamov relevantných pre bezpečnostný incident v časti výskumu týkajúcom sa názvu.

Potom sme na údaje použili jednu zo štyroch agregačných funkcií (súčet, maximum, priemer alebo medián), výsledkom čoho bolo 487 záznamov pri analýze inodov súboru a 715 záznamov pri analýze názvov súborov.

Vytvorili sme aj kombinácie atribútov, ktoré obsahovali atribút inode alebo name, na základe ktorých sme agregovali údaje. Zvažovalo sa celkovo 126 kombinácií vyššie uvedených siedmich kategórií atribútov, od použitia jednej kategórie atribútov až po použitie všetkých kategórií atribútov.

3.3 Metódy na hľadanie anomálií

Detekcia outlierov je časť výskumu, ktorá má uplatnenie v mnohých oblastiach. Používa sa na detekciu anomálií, kde máme záujem odhaliť abnormálne alebo neobvyklé pozorovania. Metódy na detekciu outlierov môžu byť s učiteľom, bez učiteľa alebo kombinované učenie (čiastočne s učiteľom a čiastočne bez učiteľa). Pri detekcii outlierov bez učiteľa pracujeme s neoznačenými údajmi. Nedokážeme určiť, ktoré záznamy z údajov sú odľahlé a ktoré nie. Hlavným cieľom je použiť metódu detekcie outlierov na identifikáciu objektov, ktoré sú odľahlé v rámci datasetu [67].

V tomto výskume sme sa zamerali na metódy na detekciu outlierov bez učiteľa, pretože počas vyšetrovania pracujeme s neoznačkovanými údajmi. Tieto metódy môžeme rozdeliť do štyroch kategórií [68]:

- **Lineárne modely** – napríklad Principal Component Analysis (PCA), One-Class Support Vector Machines using Stochastic Gradient Descent (SGDOCSVM);
- **Modely založené na blízkosti** – napríklad Local Outlier Factor (LOF), Subspace Outlier Detection (SOD), Rotation-based Outlier Detection (ROD);
- **Pravdepodobnostné modely** – napríklad Empirical-Cumulative-distribution-based Outlier Detection (ECOD), Copula Based Outlier Detection (COPOD);
- **Súbory outlierov (Outlier Ensembles)** – napríklad Isolation Forest, Isolation-based Anomaly Detection Using Nearest-Neighbor Ensembles (INNE), Lightweight on-line detector of anomalies (LODA);

Na implementáciu metód sme použili dve python knižnice: scikit-learn a pyod. Scikit-learn [69] je knižnica s otvoreným zdrojovým kódom používaná na analýzu údajov. Pozostáva z rôznych metód klasifikácie, regresie a klastrovania. Pyod [70] je knižnica na zisťovanie odľahlých objektov vo viacrozmerných údajoch. Obsahuje viac ako 40 detekčných algoritmov ako LOF alebo ECOD.

Tab. 2 Porovnanie štyroch metód bez učiteľa

Metóda	**Kategória**	**Parametre**	**Časová zložitosť**
ECOD	Pravdepodobnostné	contamination	O(n.d)
IForest	Súbory outlierov	contamination, no_estimators	O(n)
LOF	Založené na blízkosti	contamination, no_neighbors, metric	$O(n^2)$
PCA	Lineárny model	contamination	$O(p^2.n+p^3)$

Tab. 2 predstavuje porovnanie rôznych metód, kde n označuje počet atribútov, d predstavuje počet dimenzií a p označuje počet prvkov. Uvádzame aj časovú zložitosť všetkých metód, podľa čoho predpokladáme, že metódy LOF budú mať najdlhší čas

vykonávania. Skúmame tiež vplyv kontaminácie naprieč všetkými zvažovanými metódami. Kontaminácia definuje podiel odľahlých hodnôt v datasete. Okrem toho v prípade metódy IForest skúmame vplyv počtu odhadov, zatiaľ čo v prípade LOF skúmame vplyv počtu susedov a zvolenej metriky.

3.3.1 Lineárne modely (Linear models)

Ako reprezentanta tohto prístupu sme zvolili **Principal Component Analysis (PCA)**. Ide o lineárnu redukciu dimenzionality pomocou singulárneho rozkladu údajov na ich premietnutie do nižšie dimenzionálneho priestoru [71, 72]. Hlavným cieľom tejto metódy je extrahovať dôležité informácie z údajov a reprezentovať ich ako množinu nových ortogonálnych premenných nazývaných hlavné komponenty [73]. Pre túto metódu sme zvážili rôzne hodnoty kontaminácie:

- **kontaminácia** (contamination) – 0.001, 0.002, 0.005, 0.01, 0.02, 0.05 a 0.1.

3.3.2 Modely založené na blízkosti (Proximity-based models)

Ako reprezentanta tohto prístupu sme zvolili **Local Outlier Factor (LOF)**. LOF je metóda detekcie anomálií bez učiteľa. Vypočíta lokálnu odchýlku hustoty daného dátového bodu od jeho susedov [74, 75]. Metóda LOF má tri kľúčové parametre: metriku vzdialenosti, počet susedov a úroveň kontaminácie v údajoch. Metrika vzdialenosti sa používa na výpočet vzdialenosti medzi údajovými bodmi a počet susedov určuje, koľko susedných bodov sa berie do úvahy v analýze. V našej analýze sme experimentovali s rôznymi hodnotami týchto parametrov LOF:

- **metrika** (metric) – braycurtis, canberra, chebyshev, cityblock, correlation, cosine, euclidean, minkowski, sqeuclidean a jaccard,
- **počet susedov** (number of neighbors) – 2 , 5, 10, 15, 20, 30 a 50,
- **kontaminácia** (contamination) – 0.001, 0.002, 0.005, 0.01, 0.02 a 0.05.

3.3.3 Pravdepodobnostné modely (Probabilistic models)

Ako reprezentanta tohto prístupu sme zvolili **Empirical-Cumulative-distribution-based Detection Outlier Detection (ECOD)**. Je to bezparametrový, vysoko interpretovateľný algoritmus detekcie outlierov založený na empirických kumulatívnych distribučných funkciách (CDF). Funkcie CDF sa používajú na opis rozdelenia pravdepodobnosti náhodných premenných. Táto metóda je inšpirovaná skutočnosťou, že odľahlé hodnoty sú často zriedkavé alebo anomálne udalosti [76]. Keďže ECOD je metóda bez parametrov, zvažovali sme rôzne hodnoty iba pre kontamináciu:

- **kontaminácia** (contamination) – 0.001, 0.002, 0.005, 0.01, 0.02, 0.05 a 0.1.

3.3.4 Súbory outlierov (Outlier ensembles)

Ako reprezentanta posledného prístupu sme zvolili **Isolation Forest (IForest)**. Efektívnym spôsobom detekcie outlierov vo vysokorozmerných súboroch údajov je použitie náhodných lesov. Náhodné rozdelenie vytvára viditeľne kratšie cesty pre anomálie. Preto, keď IForest produkuje kratšie dĺžky ciest pre konkrétne vzorky, je veľmi pravdepodobné, že sa jedná o anomálie. Je to metóda, ktorá izoluje pozorovania náhodným výberom prvku a následným výberom delenej hodnoty medzi maximálnou a minimálnou hodnotou vybranej funkcie [77]. Pre túto metódu sme zvolili rôzne hodnoty pre množstvo odhadov, ktoré sa týkajú počtu stromov v „lese“, a kontaminácie:

- **počet stromov v lese** (number of estimators) – 10, 20, 50 a 100,
- **kontaminácia** (contamination) – 0.001, 0.002, 0.005, 0.01, 0.02 a 0.05.

3.4 Metriky na vyhodnotenie

Nasledujúce metriky skúmajú výkonnosť metód v našom výskume. Pre tento výskum používame na vyhodnotenie maticu zámeny. Metriky, ktoré sa často používajú na vyhodnotenie účinnosti metód detekcie anomálií, sú F1 skóre (F1 Score), Precision a Recall.

Recall meria schopnosť klasifikátora správne identifikovať skutočné pozitíva. Je to cenná metrika na zisťovanie všetkých TP, aj keď výsledkom je vyšší počet FP. Táto

metrika hodnotí účinnosť identifikácie všetkých anomálií a uprednostňuje minimalizáciu počtu FN. Na druhej strane **precision** odráža presnosť pozitívnych predpovedí, ktoré sú v skutočnosti pozitívne.

F1 skóre [78] je metrika používaná na vyhodnotenie modelu, pričom sa berie do úvahy Precision a Recall nasledovne:

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

kde TP je True Positives, TN je True Negatives, FP je False Positives a FN je False Negatives a:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

V rámci nášho výskumu máme celkom **487 záznamov** (pre inode) - 472 sú inlieri, 15 sú outlieri a **715 záznamov** (pre názov) - 698 sú inlieri, 17 sú outlieri. **TP** je maximálne 15 (pre inode) a 17 (pre názov). **FP** je Outl_count – Rate (pre inode a názov). **FN** je 15 – Rate (pre inode), 17 – Rate (pre názov) a **TN** je 472 – FP (pre inode), 715 – FP (pre názov).

4 Parciálne výsledky

V tejto kapitole uvádzame parciálne výsledky pre obe časti výskumu (inody súborov aj názvy súborov). Predstavujeme a diskutujeme vhodnú množinu atribútov reprezentujúcu digitálnu stopu v súborovom systéme. Poskytujeme tiež výsledky a diskutujeme o vhodnom použití agregačných funkcií a parametrov vybraných metód na detekciu outlierov.

4.1 Celkové výsledky

V Tab. 3 a Tab. 4 uvádzame celkové výsledky nášho skúmania. Pre každú metódu poskytujeme maximálne, stredné hodnoty a medián pre F1 skóre. Použitie inodov súboru poskytuje značne lepšie výsledky. Znamená to, že je vhodnejšie agregovať digitálne stopy, ktoré je možné získať zo súborového systému, podľa inodov. V týchto tabuľkách sú najlepšie výsledky zvýraznené tučným písmom. ECOD vychádza ako najlepšia metóda, keďže má najlepšie výsledky okrem maximálnej hodnoty F1 skóre pre analýzu názvov súborov. V tomto prípade dosiahla najlepší výsledok metóda IForest.

Tab. 3 F1 skóre pre analýzu inodov

Metóda	**F1 skóre max**	**F1 skóre priemer**	**F1 skóre medián**
ECOD	**0.8000**	**0.3056**	**0.2727**
IForest	**0.8000**	0.2345	0.1250
LOF	**0.8000**	0.1286	0.0952
PCA	0.7333	0.1426	0.0000

Tab. 4 F1 skóre pre analýzu názvov súborov

Metóda	**F1 skóre max**	**F1 skóre priemer**	**F1 skóre medián**
ECOD	0.5000	**0.1632**	**0.1600**

IForest	**0.5600**	0.1306	0.1053
LOF	0.5185	0.0755	0.0377
PCA	0.4800	0.1263	0.1053

V Tab. 5 uvádzame **5 najlepších výsledkov pre inode** pre každú vybranú metódu detekcie outlierov. Výsledky sme zoradili podľa F1 skóre, pretože táto metrika sa často používa na hodnotenie metód na detekciu anomálií. Keďže porovnávame metódy podľa F1 skóre, najlepšie výsledky (0,80) dosiahli tri metódy – LOF, ECOD a IForest. V prípade metódy PCA bolo najlepšie F1 skóre približne 0,733. Všeobecne najlepšie výsledky sa teda dosiahli pomocou agregačnej funkcie súčet (sum).

Tab. 5 Celkové výsledky pre analýzu inodov

Metóda	**Atribúty**	**Agg**	**TP**	**Recall**	**Precision**	**F1 skóre**
PCA	15	sum	11	0.7333	0.7333	0.7333
PCA	13	sum	9	0.6	0.9	0.72
PCA	107	median	9	0.6	0.9	0.72
PCA	42	median	9	0.6	0.9	0.72
PCA	122	median	9	0.6	0.9	0.72
LOF	53	sum	12	0.8	0.8	0.8
LOF	46	sum	10	0.6667	1.0	0.8
LOF	30	sum	10	0.6667	1.0	0.8
LOF	76	sum	10	0.6667	1.0	0.8
LOF	58	sum	10	0.6667	1.0	0.8
ECOD	111	sum	10	0.6667	1.0	0.8

ECOD	34	sum	10	0.6667	1.0	0.8
ECOD	102	sum	10	0.6667	1.0	0.8
ECOD	100	sum	10	0.6667	1.0	0.8
ECOD	119	sum	10	0.6667	1.0	0.8
IForest	86	sum	12	0.8	0.8	0.8
IForest	69	sum	10	0.6667	1.0	0.8
IForest	66	sum	10	0.6667	1.0	0.8
IForest	111	sum	10	0.6667	1.0	0.8
IForest	99	sum	10	0.6667	1.0	0.8

V Tab. 6 uvádzame **5 najlepších výsledkov pre názov** každej metódy detekcie outlierov. Najlepšie výsledky (0,56) sme dosiahli pri použití metódy IForest. Výsledky sú však vo všeobecnosti neuspokojivé, pretože skóre F1 sa pohybuje tesne nad a pod 0,50. Funkciou agregácie, ktorá dosiahla najlepšie výsledky, je funkcia max.

Tab. 6 Celkové výsledky pre analýzu názvov súborov

Metóda	**Atribúty**	**Agg**	**TP**	**Recall**	**Precision**	**F1 skóre**
PCA	45	max	6	0.3529	0.75	0.48
PCA	52	max	6	0.3529	0.75	0.48
PCA	88	max	6	0.3529	0.75	0.48
PCA	17	max	6	0.3529	0.75	0.48
PCA	27	max	6	0.3529	0.75	0.48
LOF	6	sum	7	0.4118	0.7	0.5185

LOF	25	max	7	0.4118	0.6364	0.5
LOF	25	max	7	0.4118	0.6364	0.5
LOF	25	max	7	0.4118	0.6364	0.5
LOF	25	max	7	0.4118	0.6364	0.5
ECOD	76	mean	8	0.4706	0.5333	0.5
ECOD	82	mean	8	0.4706	0.5333	0.5
ECOD	70	mean	8	0.4706	0.5333	0.5
ECOD	40	mean	8	0.4706	0.5333	0.5
ECOD	56	mean	8	0.4706	0.5333	0.5
IForest	86	max	7	0.4118	0.875	0.8
IForest	40	mean	8	0.4706	0.6154	0.8
IForest	56	max	8	0.4706	0.6154	0.8
IForest	40	max	6	0.3529	0.8571	0.5
IForest	86	max	7	0.4118	0.6364	0.5

4.2 Analýza výkonu

Čas vykonania kódu bol základným parametrom na porovnanie, pretože rýchla reakcia je kľúčová pri riadení bezpečnostných incidentov, aj keď nemusí byť úplne presná.

Tab. 7 Porovnanie výkonu štyroch vybraných metód strojového učenia

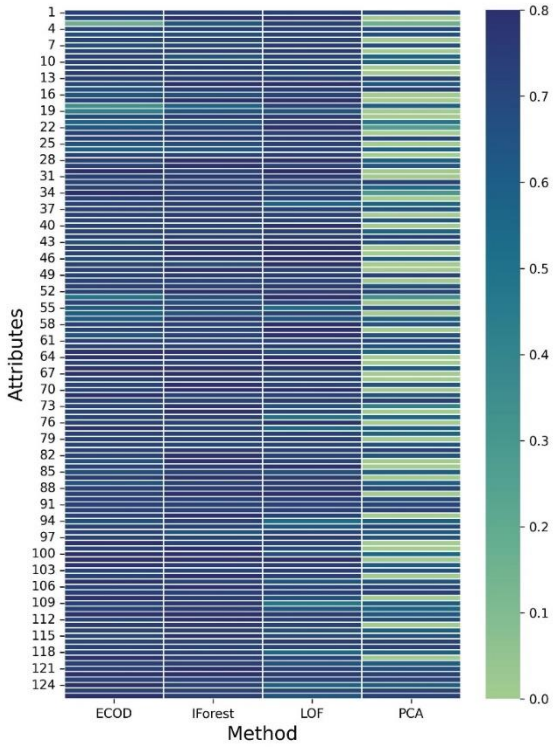
Metóda	**Počet parametrov**	**Časová zložitosť**	**Čas behu**
ECOD	1	O(n.d)	30s

IForest	2	O(n)	1439s
LOF	3	$O(n^2)$	42967s
PCA	1	$O(p^2.n + p^3)$	41s

Tab. 7 predstavuje porovnanie výkonnosti štyroch reprezentatívnych metód. Kvôli kvadratickej zložitosti metódy LOF vo vzťahu k počtu údajových bodov a zohľadneniu troch parametrov súčasne trvalo dokončenie kódu viac ako 40 000 sekúnd. Napriek lineárnej časovej zložitosti IForest, analýza dvoch parametrov viedla k času vykonaniu kódu viac ako 1 400 sekúnd. Metódy ECOD a PCA predčili ostatné, čo čiastočne súviselo aj s počtom skúmaných parametrov.

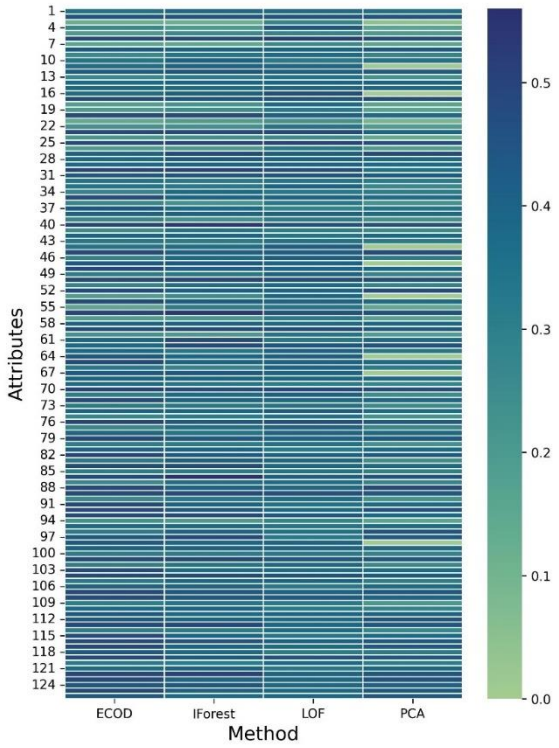
4.3 Analýza atribútov

Na Obr. 1 môžeme vidieť teplotnú mapu (heatmap) maximálneho F1 skóre pre každú metódu detekcie inodov súboru. Ako vidíme, výsledky pre PCA sú vo všeobecnosti nedostatočné - v mnohých prípadoch je hodnota 0.



Obr. 1 Teplotná mapa maximálnej hodnoty F1 skóre pre atribúty a metódy detekcie outlierov pre inody súborov

Obr. 2 zobrazuje výsledky analýzy atribútov pre názvy súborov. Ako vidíme výsledky sú podľa farieb v teplotnej mape na prvý pohľad oveľa horšie ako vo fáze analýzy atribútov pre inody.



Obr. 2 Teplotná mapa maximálnej hodnoty F1 skóre pre atribúty a metódy detekcie outlierov pre názvy súborov

Podľa Obr. 1 a Obr. 2 sme vybrali niekoľko kombinácií atribútov, aby sme demonštrovali príklad výsledkov.

V Tab. 8 uvádzame desať rôznych kombinácií atribútov pre inody. Pre atribúty zo skupiny file_type (3) môžeme vidieť, že ECOD a PCA dosiahli nedostatočné výsledky, kým LOF dosiahol dobré výsledky. Zaujímavou kombináciou je skupina atribútov timestamp, file_type, file_type2 a file_format (76). Vidíme, že pre ECOD, IForest a LOF máme hodnotu F1 skóre väčšiu ako 0,72, ale PCA dosiahla hodnotu 0.

Tab. 8 Porovnanie atribútov pri analýze inodov

Atribúty	**Asociované atribúty**	**F1 max – ECOD**	**F1 max – IForest**	**F1 max – LOF**	**F1 max – PCA**
3	file_type	0.16	0.62	0.72	0.16

15	source_type, dir_type	0.67	0.73	0.73	0.73
18	file_type, dir_type	0.34	0.58	0.72	0.59
19	file_type, file_type2	0.42	0.64	0.64	0
21	file_type, file_size	0.57	0.64	0.75	0.48
24	dir_type, file_size	0.72	0.75	0.75	0.7
53	file_type, dir_type, file_size	0.48	0.67	0.8	0.33
72	timestamp, file_type, dir_type, file_format	0.72	0.72	0.75	0.72
76	timestamp, file_type, file_type2, file_format	0.72	0.72	0.8	0
121	timestamp, source_type, file_type, dir_type, file_format, file_size	0.72	0.8	0.72	0.67

V Tab. 9 uvádzame desať rôznych kombinácií atribútov pre názvy súborov. Pre kombináciu file_type a file_size (21) vidíme, že všetky metódy vo všeobecnosti dosiahli neuspokojivé výsledky. Kombináciou source_type, file_type, file_type2 a file_format (86) dosiahla metóda IForest najlepšie výsledky (0,56).

Tab. 9 Porovnanie atribútov pri analýze názvov súborov

Atribúty	Asociované atribúty	F1 max – ECOD	F1 max – IForest	F1 max – LOF	F1 max – PCA
6	file_format	0.48	0.48	0.52	0.48
21	file_type, file_size	0.14	0.19	0.23	0.09
25	file_type2, file_format	0.48	0.5	0.5	0.47
36	timestamp, file_type, file_size	0.27	0.23	0.32	0.23
50	source_type, file_type2, file_format	0.48	0.5	0.47	0.47
56	file_type, file_type2, file_format	0.5	0.53	0.45	0.45
86	source_type, file_type, file_type2, file_format	0.48	0.56	0.41	0.45
94	file_type, dir_type, file_type2, file_size	0.23	0.19	0.29	0.16
109	timestamp, file_type, dir_type, file_type2, file_size	0.3	0.33	0.32	0.21

122	timestamp, source_type, file_type, file_type2, file_format, file_size	0.48	0.5	0.4	0.44

V Tab. 10 a Tab. 11 uvádzame vplyv atribútov na F1 skóre pre každú z vybraných metód. Hodnoty v stĺpcoch ECOD, IForest, LOF a PCA sú priemerom stredných hodnôt F1 skóre pre každú metódu detekcie outlierov a skupinu atribútov. Stĺpec 'Occur' obsahuje binárne hodnoty 0 a 1. 0 znamená, že daná skupina atribútov sa pri použití zodpovedajúceho modelu nevyskytla a teda nemala žiadny vplyv na výslednú hodnotu F1 skóre. Naopak, 1 znamená, že daná skupina atribútov sa vyskytla pri použití príslušného modelu a teda ovplyvnila výslednú hodnotu F1 skóre. Týmto spôsobom je možné porovnať vplyv danej skupiny atribútov na výsledné skóre F1.

Tab. 10 Vplyv atribútov na výsledky – inode

Atribút	**Výskyt**	**ECOD**	**IForest**	**LOF**	**PCA**
timestamp	0	0.2986	0.2333	0.1499	0.1268
timestamp	1	0.314	0.2357	0.1079	0.158
source_type	0	0.2693	0.1955	0.122	0.1324
source_type	1	0.3418	0.2735	0.1356	0.1529
file_type	0	0.3157	0.2418	0.1512	0.1444
file_type	1	0.2957	0.2274	0.1066	0.1409
dir_type	0	0.3138	0.2463	0.1563	0.1362
dir_type	1	0.2976	0.2231	0.1018	0.1489

file_type2	0	0.2996	0.2322	0.1341	0.1467
file_type2	1	0.3113	0.2367	0.1232	0.1387
file_format	0	0.2577	0.1975	0.1029	0.1329
file_format	1	0.3519	0.2704	0.1534	0.152
file_size	0	0.3012	0.2348	0.1371	0.0382
file_size	1	0.3099	0.2342	0.12	0.247

Tab. 11 Vplyv atribútov na výsledky – názov súboru

Atribút	**Výskyt**	**ECOD**	**IForest**	**LOF**	**PCA**
timestamp	0	0.1473	0.1192	0.0753	0.1202
timestamp	1	0.1785	0.1417	0.0757	0.1322
source_type	0	0.1506	0.1185	0.0671	0.1153
source_type	1	0.1758	0.1427	0.0839	0.1373
file_type	0	0.1783	0.141	0.0796	0.1379
file_type	1	0.1485	0.1206	0.0715	0.115
dir_type	0	0.1713	0.1331	0.086	0.1372
dir_type	1	0.1553	0.1282	0.0653	0.1157
file_type2	0	0.1514	0.1251	0.0707	0.1256
file_type2	1	0.1746	0.1360	0.0802	0.1269
file_format	0	0.1103	0.0849	0.0599	0.0719
file_format	1	0.2145	0.1749	0.0906	0.179
file_size	0	0.1888	0.1554	0.0883	0.1289

file_size	1	0.1376	0.1058	0.0627	0.1236

Ako môžeme vidieť, ak sa atribúty označené **timestamp** vyskytli počas behu metódy, hodnota F1 skóre bola lepšia pre takmer každú metódu - iba analýza inodov pomocou LOF dosiahla vo všeobecnosti horšie výsledky. Ak sa kombinácia atribútov označených **source_type** vyskytla počas behu, hodnota F1 skóre bola v každom prípade lepšia. Výskyt skupiny atribútov označených **file_type** vo všeobecnosti zhoršuje výsledky. Atribúty, ktoré určujú typ súboru: „filef“, „dir“ a „link“ vo všeobecnosti zhoršujú detekciu outlierov a nemali by sa používať. Výskyt **dir_type** vo väčšine prípadov zhoršuje výsledky, ale použitím metódy PCA pri analýze inodov sme dosiahli lepšie výsledky. Hodnota F1 skóre pri použití aj nepoužití **file_type2**, je veľmi podobná v oboch prípadoch pre všetky metódy. Použitie **file_format** ukazuje, že výskyt tejto kombinácie je dôležitý – vo všetkých prípadoch zlepšilo výsledky. Kombinácia označená ako **file_size** je dôležitá pre PCA v rámci analýzy inodov. Ako vidíme, keď sa nepoužije file_size, maximálna hodnota F1 skóre je približne len 0,038, ale keď sa použije, hodnota je približne 0,247.

Zaujímavosťou je, že file_size výrazne zvyšuje hodnotu F1 skóre pri analýze inodov súboru metódou PCA. Najvýraznejšie zlepšenie v analýze názvov súborov možno vidieť v metóde ECOD s file_format. Vo všeobecnosti skupiny atribútov s názvom timestamp, source_type a file_format zvyšujú hodnotu F1 skóre. Z praktického hľadiska odporúčame zvážiť tieto atribúty v modeloch na detekciu outlierov, ale aj pri klasifikácii digitálnych stôp, ktoré je možné získať zo súborového systému a pri hľadaní vzťahov medzi nimi.

4.4 Analýza agregačných funkcií

V tejto časti výskumu sme použili štyri už spomínané **agregačné funkcie** – súčet, maximum, priemer a medián (sum, max, mean, median). Pri agregačnej funkcii max je výsledkom agregácie 0, ak sa atribút nevyskytuje medzi agregovanými prvkami. Naopak, 1 je výsledkom agregácie, ak sa atribút vyskytuje aspoň raz. Ostatné agregačné funkcie sú triviálne. S použitím agregačnej funkcie súčtu sme vo všeobecnosti dosiahli najlepšie výsledky.

Tab. 12 Porovnanie výsledky pre agregačné funkcie pri analýze inodov

Agregačná funkcia	Metóda	F1 skóre max	F1 skóre priemer	F1 skóre medián
max	ECOD	0.72	0.25	0.2
	IForest	0.72	0.1701	0.125
	LOF	0.75	0.1229	0.1
	PCA	0.72	0.0925	0.0
mean	ECOD	0.72	0.3477	0.3333
	IForest	0.75	0.2384	0.16
	LOF	0.75	0.1206	0.0526
	PCA	0.72	0.1461	0.087
median	ECOD	0.75	0.2624	0.2353
	IForest	0.75	0.187	0.1250
	LOF	0.75	0.1205	0.0
	PCA	0.72	0.1361	0.0
sum	ECOD	0.8	0.3622	0.3333
	IForest	0.8	0.3425	0.3158
	LOF	0.8	0.1503	0.1143
	PCA	0.7333	0.1958	0.125

V Tab. 12 uvádzame výsledky analýzy inodov súborov pre agregačné funkcie. Pre agregačnú funkciu **max** bola dosiahnutá maximálna hodnota pre F1 skóre pomocou LOF (0,75). ECOD, IForest a PCA dosiahli hodnotu 0,72. Priemer a medián F1 skóre pri použití PCA vo všeobecnosti dosiahli nedostatočné hodnoty. Pre agregačnú funkciu

priemer (mean) najlepšie výsledky dosiahli IForest a LOF (0,75). Najlepšie hodnoty priemeru a mediánu F1 skóre boli dosiahnuté metódou ECOD. Pomocou agregačnej funkcie **medián** sme dosiahli najlepšie výsledky pre F1 skóre metódami ECOD, IForest a LOF (0,75).

Ako sme už spomenuli, najlepšie výsledky sa vo všeobecnosti dosiahli pomocou agregačnej funkcie **súčet** (sum**)**. Maximálna hodnota F1 skóre bola dosiahnutá metódami ECOD, IForest a LOF (0,80). Priemer a medián sú vo všeobecnosti podstatne vyššie pre túto agregačnú funkciu ako ostatné.

Tab. 13 Porovnanie výsledky pre agregačné funkcie pri analýze názvov súborov

Agregačná funkcia	**Metóda**	**F1 skóre max**	**F1 skóre priemer**	**F1 skóre medián**
max	ECOD	0.4828	0.2088	0.2
	IForest	0.56	0.1622	0.1538
	LOF	0.5	0.0939	0.08
	PCA	0.48	0.1752	0.16
mean	ECOD	0.5	0.19	0.1905
	IForest	0.5333	0.157	0.125
	LOF	0.4444	0.0634	0.0
	PCA	0.4667	0.123	0.0952
median	ECOD	0.4	0.1054	0.0243
	IForest	0.4138	0.1031	0.0
	LOF	0.4286	0.0648	0.0
	PCA	0.4138	0.0911	0.0

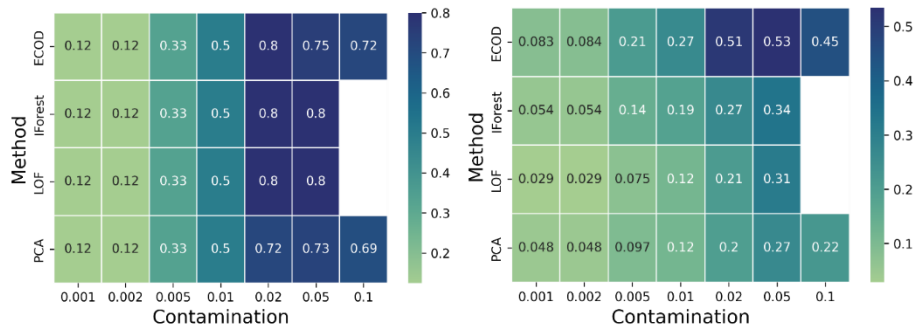
sum	ECOD	0.48	0.1485	0.1356
	IForest	0.4	0.1003	0.0952
	LOF	0.5185	0.0799	0.069
	PCA	0.3462	0.1159	0.1053

V Tab. 13 uvádzame výsledky analýzy názvov súborov pre agregačné funkcie. Ako vidíme, výsledky sú vo všeobecnosti nedostatočné. Maximálna hodnota F1 skóre sa pohybuje od približne 0,34 do 0,56 pre každú agregačnú funkciu. Najhoršie výsledky dosiahla PCA s použitím agregačnej funkcie súčet. Na druhej strane najlepšie výsledky dosiahol IForest s použitím agregačnej funkcie max.

4.5 Analýza parametra kontaminácia

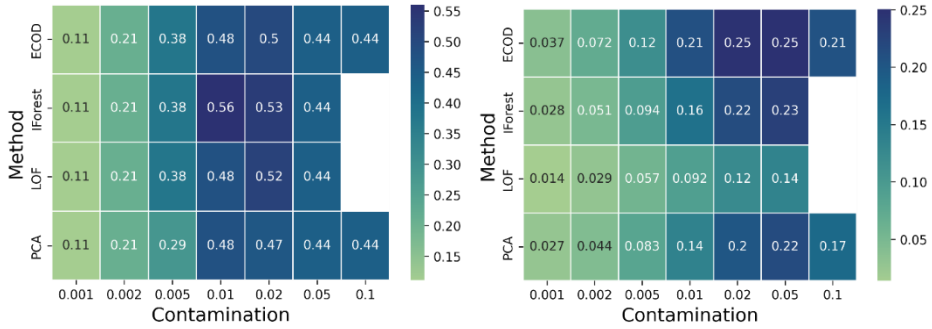
V rámci nášho výskumu sme skúšali rôzne hodnoty parametra kontaminácia pre všetky uvažované metódy. Na Obr. 3 a Obr. 4 zobrazujeme výsledky analýzy inodov a názvov súborov. Každý obrázok obsahuje dve časti. Ľavý obrázok zobrazuje teplotnú mapu maximálnych hodnôt F1 skóre pre parameter kontaminácia a metódy detekcie outlierov v prípade inodov alebo názvov súborov. Pravý obrázok zobrazuje teplotnú mapu stredných hodnôt F1 skóre pre parameter kontaminácia a metódy detekcie outlierov v prípade analýzy inodov súborov alebo názvov súborov.

Na Obr. 3 uvádzame výsledky pre inody súborov. ECOD, IForest a LOF dosiahli maximálnu hodnotu F1 skóre (0,8), zatiaľ čo kontaminácia bola nastavená na 0,02. Znamená to, že iba 2% datasetu možno považovať za outlierov. IForest a LOF dosiahli rovnaké výsledky, zatiaľ čo kontaminácia bola nastavená na 0,05. Priemerná hodnota bola najlepšia pri použití ECOD a kontaminácia bola nastavená na 0,05.



Obr. 3 Teplotné mapy maximálnych a stredných hodnôt F1 skóre pre kontamináciu a metódy detekcie outlierov pre inody súborov

Na Obr. 4 uvádzame výsledky analýzy názvov súborov. Maximálna hodnota F1 skóre (0,56) bola dosiahnutá pri použití IForest a kontaminácii nastavenej na 0,01. V prípade priemerného F1 skóre bola najvyššia dosiahnutá priemerná hodnota 0,25 pri použití ECOD, pričom kontaminácia bola nastavená na 0,02 alebo 0,05.



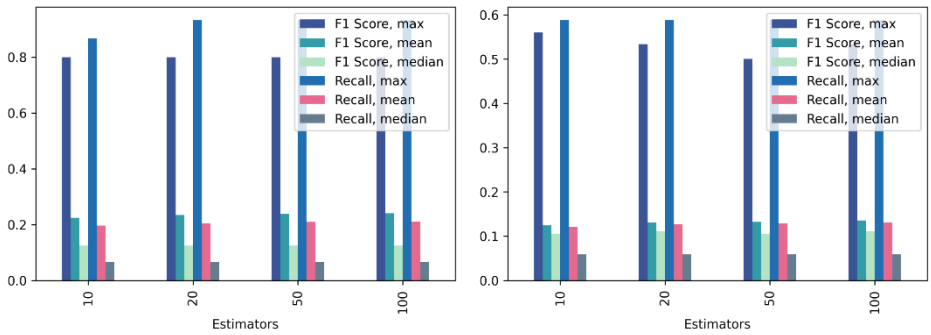
Obr. 4 Teplotné mapy maximálnych a stredných hodnôt F1 skóre pre kontamináciu a metódy detekcie outlierov pre názvy súborov

4.6 Analýza konkrétnych parametrov metód na detekciu anomálií

V tejto podkapitole sa zameriavame na špecifické parametre metód na detekciu outlierov. Metódy ECOD a PCA nemajú špecifické parametre, ak za špecifický parameter nepovažujeme kontamináciu.

Isolation Forest

V prípade metódy IForest je podstatný parameter **počet stromov** (number of estimators). Ako sme uviedli, tento parameter predstavuje počet základných stromov v súbore, ktoré budú v lese postavené. Predvolená hodnota je 100. V tomto výskume sme testovali 10, 20, 50 a 100. Na Obr. 5 môžeme vidieť maximálne, stredné hodnoty a priemer F1 skóre a Recall pre parameter metódy IForest - množstvo odhadov v prípade inodov súborov (vľavo) a názvov súborov (vpravo).



Obr. 5 Analýza parametra metódy IForest – počet odhadov

V prípade inodov súboru je maximálna hodnota F1 skóre pre všetky odhady 0,8 a stredná hodnota F1 skóre pre všetky odhady je 0,125. Stredná hodnota F1 skóre sa mierne zvyšuje so zvyšujúcim sa počtom stromov. Najvyšší počet outlierov (14) bol zaznamenaný pri nastavení na 100 stromov.

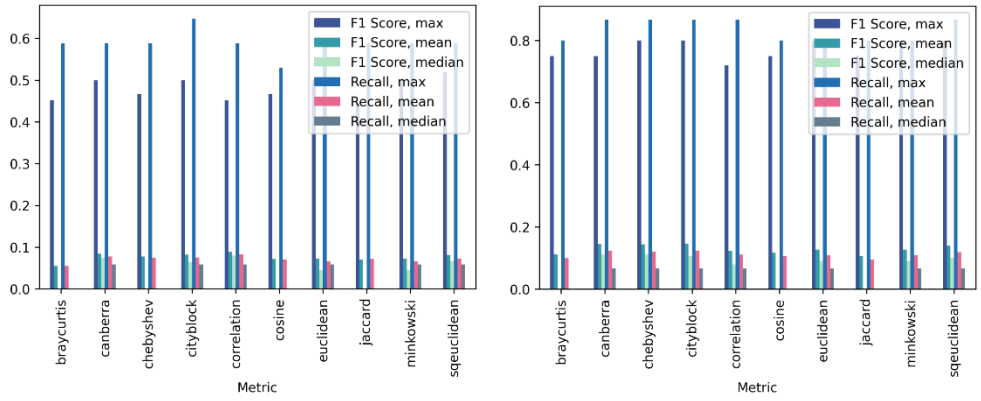
Maximálna hodnota Recall je 0,933 pre 20, 50 a 100 stromov. Priemerná hodnota Recall pre všetky počty stromov je približne 0,2. Stredná hodnota je rovnaká pre všetky počty stromov a to 0,067.

Na druhej strane v prípade názvov súborov je maximálna hodnota F1 skóre identifikovaná v prípade počtu stromov 10. Priemerná hodnota F1 skóre sa mierne zvyšuje so zvyšujúcim sa počtom stromov. Najvyšší počet identifikovaných outlierov (10) bol zaznamenaný u všetkých skúmaných počtov.

Vo všetkých prípadoch bola identifikovaná maximálna hodnota Recall (0,588). Priemerné hodnoty sa s rastúcim počtom stromov mierne zvyšujú. Hodnoty mediánu sú rovnaké pre všetky počty (0,059).

Local Outlier Factor

Najkritickejšími parametrami metódy LOF sú metrika a počet susedov. Parameter **metrika** (metric) určuje spôsob výpočtu vzdialenosti medzi dvoma susedmi v metóde LOF. V rámci výskumu sme testovali niekoľko metrík. Po aplikovaní agregačných funkcií súčtu, priemeru a mediánu dostaneme číselné hodnoty atribútov. Ďalším príkladom je agregačná funkcia max, kedy pracujeme s binárnymi dátami. Výsledné hodnoty je možné zobraziť ako binárne a číselné hodnoty. Z tohto dôvodu sme použili metriky s názvom braycurtis, canberra, chebyshev, cityblock, correlation, cosine, euclidean a sqeuclidean. Tie sú vhodné na výpočet vzdialenosti medzi dvoma číselnými vektormi. Testovali sme aj metriku nazývanú jaccard, ktorá vypočítava vzdialenosť medzi dvoma booleovskými vektormi. Nezáleží na tom, aké vektory má ako vstup (číselné alebo binárne), pretože táto metrika vie počítať aj s číselnými vektormi.



Obr. 6 Analýza parametra metódy LOF – metrika

Na Obr. 6 je možné vidieť analýzu parametra metriky v prípade inodov súboru (vľavo) a názvov súborov (vpravo). V prípade analýzy inodov súboru dosahuje maximálnu hodnotu F1 skóre niekoľko metrík (chebyshev, cityblock, euclidean, minkowski a sqeuclidean). Chebyshev sa javí byť najlepšou metrikou s dobrými maximálnymi, priemernými a strednými hodnotami pre F1 skóre.

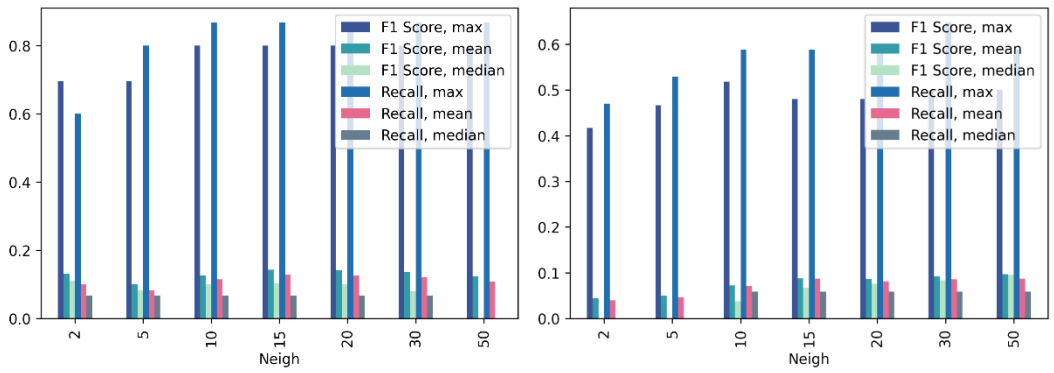
Maximálnu hodnotu Recall (0,867) dosiahli viaceré metriky (canberra, chebyshev, cityblock, correlation a sqeuclidean). Priemerná hodnota sa pohybuje od 0,095 (jaccard) do 0,124 (cityblock). Stredná hodnota bola 0,067 pre mnohé metriky, ale 0 pre braycurtis, cosine a jaccard.

V prípade analýzy názvov súborov má najlepšiu priemernú hodnotu F1 skóre metrika correlation. Najlepšia maximálna hodnota F1 skóre bola identifikovaná pomocou metriky sqeuclidean. Metrika cityblock dosiahla najvyšší počet (11) identifikovaných outlierov.

Maximálna hodnota pre Recall je 0,647 dosiahnutá metrikou nazývanou cityblock. Priemerná hodnota sa pohybuje od 0,054 (braycurtis) do 0,083 (correlation). Medián je vo väčšine prípadov 0,059 alebo 0 pre metriky braycurtis, chebyshev, cosine a jaccard.

Druhým analyzovaným parametrom v prípade metódy LOF je **počet susedov** (number of neighbors). Tento parameter určuje počet susedov, ktorí sa majú štandardne použiť pre dopyty susedov. Ak počet susedov prekročí počet poskytnutých vzoriek, použijú sa všetky vzorky. Ako je uvedené vyššie, v tomto výskume sme testovali hodnoty 2, 5, 10, 15, 20, 30 a 50.

Na Obr. 7 je možné vidieť analýzu parametra metódy LOF - počet susedov v prípade inodov súborov (vľavo) a názvov súborov (vpravo). V prípade analýzy inodov súborov sa od 10 susedov vyššie dosiahlo maximálne F1 skóre (0,8). Najvyššia priemerná hodnota F1 skóre bola identifikovaná u 15 susedov.



Obr. 7 Analýza parametra metódy LOF – počet susedov

Maximálna hodnota Recall bola 0,867 pre 10 a viac susedov. Priemerná hodnota je najvyššia pre 15 susedov (0,127). Stredná hodnota je 0,067 pre všetky prípady okrem 50 susedov.

V prípade analýzy názvov súborov sa so zvyšujúcim sa počtom susedov zvyšuje hodnota F1 skóre. Maximálna hodnota F1 skóre bola dosiahnutá v prípade 10 susedov. Pri použití 2 alebo 5 susedov sme získali neuspokojivé výsledky.

Maximálna hodnota Recall bola 0,647 pre 30 susedov. Priemerná hodnota sa pohybuje od 0,04 (2 susedia) do 0,087 (15 susedov). Stredná hodnota je 0,059 pre všetky prípady okrem 2 a 5 susedov.

4.7 Analýza outlierov

Ako sme už spomenuli, manuálne sme identifikovali pomocou digitálnej forenznej analýzy 15 inodov a 17 súborov ako digitálne stopy relevantné pre prípad.

V Tab. 14 a Tab. 15 uvádzame percento výskytov manuálne vybraných inodov alebo názvov vo výsledkoch každej metódy na detekciu outlierov. Stĺpec Inode/Názov súboru predstavuje špecifický inode súboru alebo názov súboru.

Tab. 14 Percento výskytov manuálne vybraných inodov

Inode	LOF	ECOD	PCA	IForest	Priemer
84630	8.5%	32.1%	15.4%	20.6%	19.1%
84880	13.2%	11.4%	11.2%	8.2%	11.0%
84987	28.5%	49.6%	29.6%	31.2%	34.7%
86966	0.4%	8.1%	2.9%	1.7%	3.3%
86967	3.7%	22.6%	10.1%	11.3%	11.9%
86968	22.6%	72.3%	32.9%	56.6%	46.1%
86970	11.5%	30.6%	10.2%	16.6%	17.2%
86971	12.0%	43.7%	17.5%	35.0%	27.1%
86975	9.8%	31.7%	13.1%	20.7%	18.8%

87059	0.4%	5.8%	4.8%	0.4%	2.8%
87060	8.5%	32.1%	15.4%	20.6%	19.1%
87064	8.5%	32.1%	15.4%	20.6%	19.1%
87111	3.7%	22.6%	10.1%	11.3%	11.9%
87112	14.0%	39.6%	17.5%	29.7%	25.2%
87137	22.4%	35.5%	27.2%	25.1%	27.5%
Priemer	11.2%	31.3%	15.5%	20.6%	19.7%
Všetky behy	211 680	3 528	3 528	12 096	

Tab. 15 Percento výskytov manuálne vybraných názvov súborov

Názov súboru	**LOF**	**ECOD**	**PCA**	**IForest**	**Priemer**
coreupdater.exe	17.6%	30.6%	28.5%	13.3%	22.5%
FILESH 1	13.6%	38.9%	39.1%	35.5%	31.8%
Secret	12.7%	51.3%	46.7%	36.2%	36.7%
BETH_S 1.TXT	20.7%	32.1%	25.4%	21.7%	24.9%
Beth_Secret.lnk	0.8%	0.5%	0.1%	0.0%	0.4%
SECRET 1.TXT	16.1%	42.7%	38.5%	39.9%	34.3%
SECRET_beth.lnk	0.4%	3.0%	0.5%	0.0%	1.0%
Szechuan	0.9%	14.4%	7.2%	5.4%	7.0%
SZECHU 1.TXT	16.4%	34.1%	26.9%	25.4%	25.7%
Secret.lnk	0.3%	6.9%	1.0%	0.4%	2.2%

NoJerry.lnk	0.4%	3.0%	0.5%	0.0%	1.0%
NoJerry.txt f01b4d95ca	8.5%	43.6%	23.0%	25.8%	25.2%
Destinations-ms	0.7%	5.9%	0.4%	0.5%	1.9%
SECRET_beth.txt	7.6%	5.0%	7.7%	4.3%	6.2%
Beth_Secret.txt	4.2%	13.2%	11.9%	7.8%	9.3%
Secret.zip	0.0%	0.0%	0.0%	0.0%	0.0%
coreupdater.exe.2424-urv.partial	0.0%	0.0%	0.0%	0.0%	0.0%
Priemer	7.1%	19.1%	15.1%	12.7%	
Všetky behy	211 680	3 528	3 528	12 096	

Výsledky v oboch tabuľkách ukazujú, že distribúcia nie je jednotná, pričom špecifické inody alebo názvy súborov sú častejšie identifikované ako outlieri ako iné. Hodnoty v bunkách predstavujú percento prípadov, kedy bol špecifický inode súboru alebo názov súboru identifikovaný ako outlier pomocou konkrétnej metódy na detekciu outlierov. Riadok „Všetky behy“ zobrazuje celkový počet behov konkrétnej metódy a riadok/stĺpec „Priemer“ predstavuje priemerné percento z príslušného riadka alebo stĺpca.

Ako môžeme vidieť v Tab. 14, najmenejkrát určené ako outlieri boli inody 86966 a 87059. Inode 86966 je spojený s adresárom „\FileShare\Secret“ a inode 87059 je spojený so súborom „\FileShare\Secret\Szechuan Sauce.txt“. Inode 86967 („\FileShare\Secret\NoJerry.txt“) a inode 87111 („\FileShare\Secret\Beth_Secret.txt“) majú tiež slabú detekciu. Pre tieto súbory existuje špecifický atribút dir_other s hodnotou 1. Tieto inody súborov majú rovnakú úspešnosť detekcie, pretože majú identické atribúty.

Najviac boli detegované inody 84987 a 86968, ktoré obsahovali odkazy na súbory na inode 86966 („\FileShare\Secret“), ktorý je tiež outlierom. Tieto inody majú atribút file_entry_shell_item s hodnotou 1.

Ďalším priemerne identifikovaným outlierom inode 87137, prepojený s „\Windows\System32\coreupdater.exe“, čo je malvér. Má špecifickú kombináciu atribútov dir_win a file_executable s hodnotou 1.

Pre všetkých outlierov je špecifické, že hodnota atribútov M, A, C a B je 1. Digitálne stopy s file_stat, file_entry_shell_item a filef hodnotami 1 sú častejšie identifikované.

V prípade analýzy názvov súborov môžeme v Tab. 15 vidieť, že súbory "Secret.zip" a "coreupdater.exe.2424urv.partial" neboli nikdy identifikované ako outlieri ani jednou z metód.

Názov súboru „Secret“ bol najčastejšie identifikovaný ako outlier pomocou metód ECOD (1811-krát) a PCA (1648-krát). IForest identifikoval ako outliera najčastejšie názov súboru „SECRET 1.TXT“ (4823-krát). Názov súboru „BETH_S 1.TXT“ často identifikuje LOF ako outliera. Namiesto identifikácie odkazov na súbory (súbory .lnk) ako outlierov sa identifikujú konkrétne súbory.

Vo fáze spracovania datasetu sme identifikovali dve najpočetnejšie skupiny inodov: inode 0 a inode 84656. Ako bolo uvedené vyššie v časti venovanej spracovaniu datasetu, ide o metasúbory $MFT a $UsnJrnl. V rámci datasetu bolo identifikovaných 923 agregovaných záznamov týkajúcich sa $MFT a 8272 záznamov súvisiacich s $UsnJrnl. Keďže cieľom výskumu nebolo identifikovať tieto súbory ako outlierov a zároveň by počet týchto súborov (ostatné záznamy boli počítané v jednotkách) skresľoval výsledky analýzy, rozhodli sme sa tieto záznamy vynechať. V opačnom prípade by ich každé spustenie rôznych metód označilo za outlierov. Toto opomenutie však ovplyvnilo aj detekciu súboru „Secret.zip“. Vzhľadom na vyššie uvedené zistenia je potrebné v budúcnosti zvážiť špeciálne spracovanie metasúborov $MFT a $UsnJrnl – pridať ich do pôvodného datasetu, ale vynechať ich ako detegovaných outlierov.

Hoci analýza názvov súborov vo všeobecnosti poskytuje horšie výsledky ako analýza založená na inodoch súborov, výsledky uvedené v Tab. 15 ukazujú, že niektoré súbory je možné zistiť viacerými metódami (napr. „Secret“, „SECRET~1.TXT“).

Záver

S narastajúcim množstvom kybernetických útokov a bezpečnostných incidentov sa zvyšuje aj potreba rýchlej reakcie na ne. Každý incident však môže generovať rôzne digitálne stopy, ktoré môžu, ale nemusia byť relevantné pre vyšetrovaný prípad. Z toho dôvodu sa výskumníci snažia implementovať automatizáciu do rôznych procesov v oblasti digitálnej forenznej analýzy.

V tejto práci sme sa venovali pojmom ako digitálna forenzná analýza a digitálna stopa a tiež sme popísali fázy digitálnej forenznej analýzy a výzvy, ktoré z nich pramenia. Opísali sme dataset, ktorý nám poslúžil na skúmanie metód strojového učenia a tiež použitú metodológiu. Taktiež sme popísali parciálne výsledky, ktoré sme dosiahli aplikovaním rôznych metód a variáciou ich parametrov.

Vo všeobecnosti sme v tejto práci analyzovali digitálne stopy pomocou metód strojového učenia, pričom sme sa zamerali na metódy bez učiteľa. Naša analýza bola rozdelená na dve časti: analýza inodov súborov a analýza názvov súborov. Na identifikáciu anomálií (v našom prípade relevantných digitálnych stôp) sme použili štyri metódy detekcie outlierov (ECOD, IForest, LOF a PCA). Otestovali sme rôzne hodnoty pre parametre týchto metód a sledovali sme, ako tieto hodnoty ovplyvnili výsledky. Pri analýze inodov, maximálne dosiahnuté F1 skóre bolo 0.80 použitím metód ECOD, LOF a IForest. Najlepšiu hodnotu pre medián F1 skóre (0.2727) aj priemerné F1 skóre (0.3056) dosiahla metóda ECOD. Pri analýze názvov súborov opäť dosiahol najlepšie výsledky ECOD pre medián F1 skóre (0.1600), ale tiež priemerné F1 skóre (0.1632). IForest dosiahol najvyššiu maximálnu hodnotu F1 skóre a to 0.5600. Z vyššie uvedených hodnôt vidieť, že vo všeobecnosti sme dosiahli lepšie výsledky pri analýze inodov súborov. Zhrnutím však môžeme tvrdiť, že najlepšie výsledky sme dosiahli pomocou metódy ECOD.

Ďalej sa budeme venovať vytvoreniu kvalitného, scenárovo založeného, syntetického datasetu, ktorý by spĺňal požiadavky nášho výskumu. Takisto chceme otestovať ďalšie metódy strojového učenia nie len na identifikáciu relevantných digitálnych stôp, ale tiež na hľadanie vzťahov medzi nimi a ich atribútmi.

Zoznam použitej literatúry

1. Studiawan, Hudan, and Ferdous Sohel. "Anomaly detection in a forensic timeline with deep autoencoders." Journal of Information Security and Applications 63 (2021): 103002.
2. Boddington, Richard. Practical digital forensics. Packt Publishing Ltd, 2016., Sammons, John. The basics of digital forensics: the primer for getting started in digital forensics. Elsevier, 2012.
3. Digitálna stopa [online]. [cit. 2023-08-20]. Dostupné z: https://www.swgde.org/
4. Reddy, Niranjan. Practical cyber forensics. Apress, 2019.
5. Johansen, Gerard. Digital forensics and incident response. Packt Publishing Ltd, 2017.
6. RFC 3227 [online]. [cit. 2023-08-20]. Dostupné z: https://www.ietf.org/rfc/rfc3227.txt
7. Stroeh, Kleber, Edmundo Roberto Mauro Madeira, and Siome Klein Goldenstein. "An approach to the correlation of security events based on machine learning techniques." Journal of Internet Services and Applications 4 (2013): 1-16.
8. Limmer, Tobias, and Falko Dressler. "Survey of event correlation techniques for attack detection in early warning systems." University of Erlangen, Dept. of Computer Science, Technical Report (2008): 1-37.
9. Marturana, Fabio, and Simone Tacconi. "A Machine Learning-based Triage methodology for automated categorization of digital media." Digital Investigation 10.2 (2013): 193-204.
10. Du, Xiaoyu, and Mark Scanlon. "Methodology for the automated metadata-based classification of incriminating digital forensic artefacts." Proceedings of the 14th International Conference on Availability, Reliability and Security. 2019.
11. Panchal, Esan P., Shruti B. Yagnik, and B. K. Sharma. "Use of Machine Learning Algorithm on File Metadata for Digital Forensic Investigation Process." Third International Congress on Information and Communication Technology: ICICT 2018, London. Springer Singapore, 2019.

12. Log2timeline (plaso) [online]. [cit. 2023-08-20]. Dostupné z: https://plaso.readthedocs.io/

13. Guðjónsson, Kristinn. "Mastering the super timeline with log2timeline." SANS Institute (2010).

14. Mohammad, Rami Mustafa A., and Mohammed Alqahtani. "A comparison of machine learning techniques for file system forensics analysis." Journal of Information Security and Applications 46 (2019): 53-61.

15. Studiawan, Hudan, Ferdous Sohel, and Christian Payne. "Sentiment analysis in a forensic timeline with deep learning." IEEE Access 8 (2020): 60664-60675.

16. Grajeda, Cinthya, Frank Breitinger, and Ibrahim Baggili. "Availability of datasets for digital forensics–and what is missing." Digital Investigation 22 (2017): S94-S105.

17. Luciano, Laoise, et al. "Digital forensics in the next five years." Proceedings of the 13th International Conference on Availability, Reliability and Security. 2018.

18. Casey, Eoghan. "The chequered past and risky future of digital forensics." Australian journal of forensic sciences 51.6 (2019): 649-664.

19. Mavroeidis, Vasileios, and Siri Bromander. "Cyber threat intelligence model: an evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence." 2017 European Intelligence and Security Informatics Conference (EISIC). IEEE, 2017.

20. Casey, Eoghan, et al. "Advancing coordinated cyber-investigations and tool interoperability using a community developed specification language." Digital investigation 22 (2017): 14-45.

21. Horsman, Graeme, and James R. Lyle. "Dataset construction challenges for digital forensics." Forensic Science International: Digital Investigation 38 (2021): 301264.

22. Breitinger, Frank, and Alexandre Jotterand. "Sharing datasets for Digital Forensic: a novel taxonomy and legal concerns." Forensic Science International: Digital Investigation 45 (2023): 301562.

23. Moch, Christian, and Felix C. Freiling. "The forensic image generator generator (forensig2)." 2009 Fifth International Conference on IT Security Incident Management and IT Forensics. IEEE, 2009.

24. Moch, Christian, and Felix C. Freiling. "Evaluating the forensic image generator generator." Digital Forensics and Cyber Crime: Third International ICST Conference, ICDF2C 2011, Dublin, Ireland, October 26-28, 2011, Revised Selected Papers 3. Springer Berlin Heidelberg, 2012.

25. Visti, Hannu, Sean Tohill, and Paul Douglas. "Automatic creation of computer forensic test images." Computational Forensics: 5th International Workshop, IWCF 2012, Tsukuba, Japan, November 11, 2012 and 6th International Workshop, IWCF 2014, Stockholm, Sweden, August 24, 2014, Revised Selected Papers. Springer International Publishing, 2015.

26. Scanlon, Mark, Xiaoyu Du, and David Lillis. "EviPlant: An efficient digital forensic challenge creation, manipulation and distribution solution." Digital Investigation 20 (2017): S29-S36.

27. Du, Xiaoyu, et al. "TraceGen: User activity emulation for digital forensic test image generation." Forensic Science International: Digital Investigation 38 (2021): 301133.

28. Göbel, Thomas, et al. "ForTrace-A holistic forensic data set synthesis framework." Forensic Science International: Digital Investigation 40 (2022): 301344.

29. 52 53 Sokol, Pavol, et al. "The analysis of digital evidence by Formal concept analysis." The 16th International Conference on Concept Lattices and Their Applications (CLA 2022). 2022.

30. Sokol, Pavol, et al. "Formal concept analysis approach to understand digital evidence relationships." International Journal of Approximate Reasoning 159 (2023): 108940.

31. Skopik, Florian, Markus Wurzenberger, and Max Landauer. "Detecting Unknown Cyber Security Attacks Through System Behavior Analysis." Cybersecurity of Digital Service Chains: Challenges, Methodologies, and Tools. Cham: Springer International Publishing, 2022. 103-119.

32. Beebe, Nicole, Lishu Liu, and Zi Ye. "Insider Threat Detection Using Time-Series-Based Raw Disk Forensic Analysis." Advances in Digital Forensics XIII: 13th IFIP

WG 11.9 International Conference, Orlando, FL, USA, January 30-February 1, 2017, Revised Selected Papers 13. Springer International Publishing, 2017.

33. Xu, Wei, et al. "Detecting large-scale system problems by mining console logs." Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles. 2009.

34. Studiawan, Hudan, Ferdous Sohel, and Christian Payne. "A survey on forensic investigation of operating system logs." Digital Investigation 29 (2019): 1-20.

35. Studiawan, Hudan, Christian Payne, and Ferdous Sohel. "Graph clustering and anomaly detection of access control log for forensic purposes." Digital Investigation 21 (2017): 76-87.

36. Studiawan, Hudan, and Ferdous Sohel. "Performance evaluation of anomaly detection in imbalanced system log data." 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4). IEEE, 2020.

37. Liu, Liu, et al. "Anomaly-based insider threat detection using deep autoencoders." 2018 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2018.

38. Yuan, Lun-Pin, et al. "Time-window based group-behavior supported method for accurate detection of anomalous users." 2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). IEEE, 2021.

39. Hu, Qiaona, Baoming Tang, and Derek Lin. "Anomalous user activity detection in enterprise multi-source logs." 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2017.

40. He, Shilin, et al. "Experience report: System log analysis for anomaly detection." 2016 IEEE 27th international symposium on software reliability engineering (ISSRE). IEEE, 2016.

41. Hirakawa, Rin, et al. "Large scale log anomaly detection via spatial pooling." Cognitive Robotics 1 (2021): 188-196.

42. Carrier, Brian D., and Eugene H. Spafford. "Automated Digital Evidence Target Definition Using Outlier Analysis and Existing Evidence." DFRWS. 2005.

43. Pirker, Martin, Patrick Kochberger, and Stefan Schwandter. "Behavioural comparison of systems for anomaly detection." Proceedings of the 13th International Conference on Availability, Reliability and Security. 2018.

44. Du, Xiaoyu, Quan Le, and Mark Scanlon. "Automated artefact relevancy determination from artefact metadata and associated timeline events." 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). IEEE, 2020.

45. Stolfo, Salvatore J., et al. "A comparative evaluation of two algorithms for windows registry anomaly detection." Journal of Computer Security 13.4 (2005): 659-693.

46. Tajoddin, Asghar, and Mahdi Abadi. "RAMD: registry-based anomaly malware detection using one-class ensemble classifiers." Applied Intelligence 49 (2019): 2641-2658.

47. Chouhan, Ajay S., et al. "An Ensemble Approach for Modeling Process Behavior and Anomaly Detection." Proceedings of Emerging Trends and Technologies on Intelligent Systems: ETTIS 2021. Springer Singapore, 2022.

48. Alvarez, Maxime, et al. "A revealing large-scale evaluation of unsupervised anomaly detection algorithms." arXiv preprint arXiv:2204.09825 (2022).

49. Fourure, Damien, et al. "Anomaly detection: How to artificially increase your f1-score with a biased evaluation protocol." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing, 2021.

50. Goldstein, Markus, and Seiichi Uchida. "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data." PloS one 11.4 (2016): e0152173.

51. Zoppi, Tommaso, Andrea Ceccarelli, and Andrea Bondavalli. "Evaluation of Anomaly Detection algorithms made easy with RELOAD." 2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE). IEEE, 2019.

52. Elmrabit, Nebrase, et al. "Evaluation of machine learning algorithms for anomaly detection." 2020 international conference on cyber security and protection of digital services (cyber security). IEEE, 2020.

53. Marková, Eva, Pavol Sokol, and Kristína Kováćová. "Detection of relevant digital evidence in the forensic timelines." 2022 14th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE, 2022.

54. Sari, Sari Ali, and Kamaruddin Malik Mohamad. "A review of graph theoretic and weightage techniques in file carving." Journal of Physics: Conference Series. Vol. 1529. No. 5. IOP Publishing, 2020.

55. Adderley, Nikolai, and Gilbert Peterson. "Interactive temporal digital forensic event analysis." Advances in Digital Forensics XVI: 16th IFIP WG 11.9 International Conference, New Delhi, India, January 6–8, 2020, Revised Selected Papers 16. Springer International Publishing, 2020.

56. Henseler, Hans, and Jessica Hyde. "Technology Assisted Analysis of Timeline and Connections in Digital Forensic Investigations." LegalAIIA@ ICAIL. 2019.

57. Peng, Liwen, Xiaolin Zhu, and Peng Zhang. "A machine learning-based framework for mobile forensics." 2020 IEEE 20th International Conference on Communication Technology (ICCT). IEEE, 2020.

58. Adam, Iliyasu Yahaya, and Cihan Varol. "Intelligence in digital forensics process." 2020 8th International Symposium on Digital Forensics and Security (ISDFS). IEEE, 2020.

59. Wang, Ai, and Xuedong Gao. "A variable scale case-based reasoning method for evidence location in digital forensics." Future Generation Computer Systems 122 (2021): 209-219.

60. Senkyire, Isaac Baffour, and Quist-Aphetsi Kester. "Social Engineering Cybercrime Evidence Analysis Using Formal Concept Analysis." 2021 International Conference on Cyber Security and Internet of Things (ICSIoT). IEEE, 2021.

61. Maluleke, Lethabo. "A Formal Concept Analysis Driven Ontology for ICS Cyberthreats." SACAIR 2020 (2020): 247.

62. Waziri, Victor Onomza, Abdullahi Umar, and M. O. R. U. F. U. Olalere. "E-fraud forensics investigation techniques with formal concept analysis." International Journal of Cyber-Security and Digital Forensics 3.4 (2014): 235-245.

63. Marková, Eva, et al. "Classification of malicious emails." 2019 IEEE 15th International Scientific Conference on Informatics. IEEE, 2019.

64. Marková, Eva, et al. "Malicious Emails Classification Based on Machine Learning." Proceedings of the Computational Methods in Systems and Software. Cham: Springer International Publishing, 2021. 797-810.

65. Johnson, Richard Arnold, and Dean W. Wichern. "Applied multivariate statistical analysis." (2002).

66. Case 001 – The Stolen Szechuan Sauce [online]. [cit. 2023-08-21]. Dostupné z: https://dfirmadness.com/the-stolen-szechuan-sauce/

67. Su, Xiaogang, and Chih-Ling Tsai. "Outlier detection." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1.3 (2011): 261-268.

68. Aggarwal, Charu C., and Charu C. Aggarwal. An introduction to outlier analysis. Springer International Publishing, 2017.

69. Scikit-learn knižnica [online] 24.08.2023 Dostupné z: https://scikit-learn.org/

70. Zhao, Yue, Zain Nasrullah, and Zheng Li. "Pyod: A python toolbox for scalable outlier detection." arXiv preprint arXiv:1901.01588 (2019).

71. Shyu, Mei-Ling, et al. "A novel anomaly detection scheme based on principal component classifier." Proceedings of the IEEE foundations and new directions of data mining workshop. IEEE Press, 2003.

72. Cheng, Zhangyu, Chengming Zou, and Jianwei Dong. "Outlier detection using isolation forest and local outlier factor." Proceedings of the conference on research in adaptive and convergent systems. 2019.

73. Abdi, Hervé, and Lynne J. Williams. "Principal Component Analysis.(2010)." Computational Statistics, John Wiley and Sons (2010): 433-459.

74. Breunig, Markus M., et al. "LOF: identifying density-based local outliers." Proceedings of the 2000 ACM SIGMOD international conference on Management of data. 2000.

75. Abhaya, Abhaya, and Bidyut Kr Patra. "RDPOD: an unsupervised approach for outlier detection." Neural Computing and Applications 34.2 (2022): 1065-1077.

76. Li, Zheng, et al. "Ecod: Unsupervised outlier detection using empirical cumulative distribution functions." IEEE Transactions on Knowledge and Data Engineering (2022).

77. Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." 2008 eighth ieee international conference on data mining. IEEE, 2008.

78. Sasaki, Yutaka. "The truth of the F-measure." Teach tutor mater 1.5 (2007): 1-5.

Publikácie autora rigoróznej práce

1. Marková, Eva, et al. "Classification of malicious emails." 2019 IEEE 15th International Scientific Conference on Informatics. IEEE, 2019. (Scopus, WoS)
2. Marková, Eva, et al. "Malicious Emails Classification Based on Machine Learning." Proceedings of the Computational Methods in Systems and Software. Cham: Springer International Publishing, 2021. 797-810. (Scopus)
3. Marková, Eva, Pavol Sokol, and Kristína Kováčová. "Detection of relevant digital evidence in the forensic timelines." 2022 14th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE, 2022. (Scopus)
4. Sokol, Pavol, Marková, Eva et al. "The analysis of digital evidence by Formal concept analysis." The 16th International Conference on Concept Lattices and Their Applications (CLA 2022). 2022. (Scopus)
5. Sokol, Pavol, Marková, Eva et al. "Formal concept analysis approach to understand digital evidence relationships." International Journal of Approximate Reasoning 159 (2023): 108940. (Scopus, WoS)