ТЕХНИЧЕСКИ УНИВЕРСИТЕТ–СОФИЯ

Факултет „Компютърни системи и технологии“
Катедра „Компютърни системи“

маг. инж. Гроздан Костадинов Христов

Методи и алгоритми за машинно обучение в киберсигурността

А В Т О Р Е Ф Е Р А Т

на дисертация за придобиване на образователна и научна степен

„Доктор“

Област на висше образование: 5. Технически науки

Професионално направление: 5.3. Комуникационна и компютърна техника

Научна специалност: „Системи с изкуствен интелект“

Научни ръководители:

доц. д-р инж. Аделина Алексиева-Петрова

доц. д-р инж. Иван Станков

София, 2025 г.

Дисертационният труд е обсъден и насочена за защита от Катедрения съвет на катедра „Компютърни системи“ към факултет „Компютърни системи и технологии“ на Технически университет – София на редовно заседание, проведено на 20.10.2025 г.

Публичната защита на дисертационния труд е насрочена за 27.01.2026 г. от 13:00 ч. в Конферентната зала на Библютечно – информационния център на Техническия университет – София на открито заседание на научно жури, определено със заповед № ОЖ-5.3-58 от 30.10.2025 г. на ректора на Техническия университет – София в състав:

1. проф.д-р Даниела Асенова Гоцева – председател
2. проф.д-р Милена Кирилова Лазарова-Мицева – научен секретар
3. проф.д-р Нина Василевна Синягина – член
4. проф.д-р Емил Иванов Йончев – член
5. проф.д-р Александър Богданов Бекярски – член

Рецензенти:

1. проф. д-р инж. Милена Кирилова Лазавора-Мицева
2. проф. д-р Нина Василевна Синягина

Материалите по защита са на разположение на интересуващите се в канцеларията на факултет „Компютърни системи и технологии“ на Технически университет – София, блок № 1, кабинет № 1443-А.

Дисертантът е задочен докторант към катедра „Компютърни системи“ на факултет „Компютърни системи и технологии“. Изследванията от по дисертационната разработка са направени от автора.

Автор: маг. инж. Гроздан Христов

Заглавие: Методи и алгоритми за машинно обучение в киберсигурността

Тираж: 30 броя

Отпечатано в ИПК на Техническия универсисет – София

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Актуалност на проблема

С широката употреба на технологични решения от сферата на комуникационните и информационни технологии от множество сектори на обществения живот и държавни отношения, възможността за класифициране на установени зловредни дейности и последващото им ограничаване се явява от особена важност за поддържането на нормалното функциониране на ежедневните процеси. Основна тежест от гореописаното попада върху експертите по киберсигурност, които са натоварени със задачата да анализират и класифицират настъпилите инциденти спрямо комуникационни и информационни системи, с цел поправяне на пропуските в киберсигурността, довели до успешното експлоатиране на уязвимостите в системите от страна на противника. Потенциалната възможност за автоматизиране на посочения процес или части от него се явява от съществена важност за оптимизиране на работата на експертите по киберсигурност и повишаване на ефективността на тяхната работа, което от своя страна би ускорило процеса по адресиране на открити експлоатирани уязвимости в комуникационните и информационните системи и биха затруднили бъдещата употреба на зловредни софтуерни продукти и методи от вече установения тип, което ще забави и повиши сложността на противникови хакерски групи/елементи в дейностите им, насочени срещу сигурността и целостта на разнородни цифрови системи.

Предлагането и разработването на ефективни подходи с използване на изкуствен интелект за конкретния клас задачи изисква детайлно и многопластово разбиране на съответната проблемна област и аспектите на сигурността в киберпространството. Тези подходи ще бъдат базирани както на литературните източници, анализирани в първата глава, така и на резултатите от проведените изследвания и експерименти, описани в следващите части на дисертационния труд.

Цел на дисертационния труд, основни задачи и методи за изследване

Цел на дисертационния труд е да се проучат различни методи, алгоритми и подходи от изкуствения интелект за класифициране на зловредни файлове към съответните хакерски групи/индивиди, свързани с пробиви в киберсигурността. В следствие на направеното проучване да бъде предложено софтуерно решение, което

да използва метод от сферата на изкуствения интелект за класифициране на зловредни файлове, което да улесни техния първичен и вторичен анализ.

Обект на изследване

Обект на изследване на дисертационния труд е употребата на различни алгоритми от областта на изкуствения интелект в контекста на киберсигурността, техните положителни и отрицателни страни и оптимизиране.

Предмет на изследване

Предмет на изследването е ефективното използване на концепции от класификационните алгоритми в софтуерни приложения, осигуряващи киберзащитата на комуникационни и информационни системи.

Задачи

1. Да се проучи и изследва възможността за прилагане на методи на машинното обучение при анализа на инциденти с киберсигурността, като основната функция на тези методи може да бъде решаване на задача за класификация, която да е в помощ на експертите по киберсигурност в присвояване на зловредни файлове и методология на работа към различни хакерски групи;
2. Да се намерят характеристики, които адекватно да отразяват обективните зависимости от класификационния статус;
3. Да се предложи и оптимизира модел за решаване на класификационната задача и да се приложи този модел за определяне на класификационния статус на анализираните артефакти;
4. Да се направи експериментален анализ за приложение на предложения модел, на базата на който да се преценят неговите възможности.

Практическа приложимост

Основното приложение на разглежданите методи и алгоритми е в анализа и последващо присвояване (атрибуцията) на зловреден софтуер към хакерски групи, в това число и такива, които са спонсорирани на държавно ниво.

От една страна, приложимостта на предложеното решение е в дейностите по реакцията на кибернападения, конкретно в частта анализ на използваните зловредни програмни инструкции. За целта се използват статични характеристики, извлечени от

пробите на зловредния софтуер, включително криптографски хеш стойности (MD5, SHA1, SHA256) и друга информация, за изграждане на два отделни класификатора от тип Random Forest. Чрез фокусиране върху статичния анализ и използване на интерпретируеми техники за машинно обучение, този подход предлага ценен инструмент на анализаторите по киберсигурност, подпомагайки бързото идентифициране на заплахи и подобряване на възможностите за разузнаването им.

От друга страна възможна интеграция в системи за киберзащита биха могли да се окажат изключително полезни за екипите по киберсигурност, когато се изправят пред предизвикателствата, свързани с анализа и противодействието на усъвършенствани кибератаки.

Апробация

Основните резултати от проведените изследвания в дисертационния труд са докладвани и обсъдени в серия от международни конференции и участие в един научноизследователски проект. Основните резултати са представени на научни конференции „International Scientific Conference Computer Science'2020“, „Youth Scientific Conference "Machines Innovation and Technology" 2020“, „2023 International Scientific Conference on Computer Science (COMSCI)“, „2023 31st National Conference with International Participation (TELECOM)“ и научното списание (journal) „International Journal of Innovative Research and Scientific Studies vol. 8 (1) 2025“. В референтната база данни Scopus са индексирани 3 от научните статии, като една от тях е получила международно цитиране в „Suraeva, M. & Afonasyev, M. & Kucheryavenko, D.. (2022). The Impact of Digitalization on Innovative Approaches to Economic Security in Regions. 10.1007/978-3-030-83175-2_43.“, издадена и публикувана в книгата „Digital Technologies in the New Socio-Economic Reality (pp.337-344)“.

Публикации

Основните постижения и резултати от дисертационния труд са представени на международни конференции и има една публикация в специализирано научно списание. В референтната база данни Scopus са индексирани 3 от научните статии, като една от тях е получила международно цитиране друга публикация, издадена и

публикувана в книга. Представени са 5 публикации, на 3 от които докторантът е първи автор и има една самостоятелна публикация.

Структура и обем на дисертационния труд

Настоящият дисертационен труд е в обем от 148 страници и съдържа уводна част, четири глави, заключение, списък с приноси, списък с публикации по дисертационния труд, списък с участия в научно-изследователски проекти, списък на фигури, списък на таблици, списък с използвани съкращения и библиографска справка на използваните литературни източници. Към дисертационния труд има едно приложение, съдържащо 14 страници. Трудът съдържа 8 таблици и 21 фигури. Цитирани са 111 източника.

II. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

Първа глава. Изследване и анализ на тенденции и заплахи в киберпространството и взаимовръзката им със сферата на изкуствения интелект.

Кибератаки – определение и типове

Нарастващото навлизане и употреба на технологични решения в човешкото ежедневие налага отделянето на по-голямо внимание към кибератаките и присъствието в цифровото пространство и киберсигурността като цяло. Въпреки това няма една обща дефиниция за това какво представлява една кибератака. Най – общо казано кибернападението включва всякакъв вид злонамерена дейност към информационни системи и/или данни чрез киберпространството. По отношение на типовете кибератаки една от най-често използваните класификации ги разделя на уеб-базирани и системни нападения, като всяка подсфера от които използва различни методи за компрометиране на цифрови системи и данни. Графично представяна на посочената йерархичност е илюстрирана на Фигура 1.

Фигура 1. Основни видове кибератаки - обща диаграма

В зависимост от тяхната дейност, кибернападенията могат да бъдат активни или пасивни. Целта на активна атака е да промени по специфичен начин системните ресурси на целта или да повлияе по някакъв начин на тяхната работа, докато целта на

пасивна атака е да се опита да проучи и използва информация от жертвата без да променя ресурсите ѝ.

Изкуствен интелект – определение и ключови аспекти

За да се разберат в по – голяма яснота основните аспекти и идеологията на изкуствения интелект, трябва първо да бъде дадена дефиниция какво представлява интелигентността. По – широка дефиниция за интелигентност може да бъде дадена като способност за логика, абстракция, разбиране, критично мислене, креативност и решаване на проблеми. Човешката интелигентност представлява умственото качество, което включва способности за учене от опит, адаптиране към нови обстоятелства, разбиране и управление на интелектуални идеи и прилагане на знание, за да се влияе на средата. От горепосоченото може да бъде направен извод, че в основната си дефиницията интелигентността не може да е сама по себе си когнитивна или умствена функция, а по-скоро целенасочено съчетание от различни дейности, насочени към успешна адаптация.

Основната идея зад изкуствения интелект е по някакъв начин да се моделира или създаде интелигентност в това, което повечето хора възприемат като неживи обекти, оттам и основният фокус върху начините, по които компютърът може да симулира и проявява интелигентно мислене. Мнозина считат, че голяма стъпка в развитието и изследването на ИИ е теорията на Алън Тюринг, публикувана в "Computing Machinery and Intelligence". В тази теория Тюринг разглежда проблема за определяне дали нещо е интелигентно или не. За решаване на проблема е предложена игра, в която целта е да се определи кой участник е истинският човек, кой е машината и ако не може да се даде еднозначен отговор или машината бъде избрана за истински човек, тогава машината преминава теста и се счита за интелигентна.

Въпреки че теорията на Алън Тюринг е по-стара от 70 години и са постигнати множество напредъци в областта на изкуствения интелигент, все още продължават спорове дали тестът на Тюринг е достатъчен за класифициране на машина като интелигентна. В статията си "Subcognition and the limits of the Turing Test" авторът Робърт М. Френч се аргументира, че като истински тест за интелигентност вариантът на Тюринг е в същност безполезен поради способността си да прониква в най-основните и най-дълбоки аспекти на човешкия интелект. В статията се подчертава, че

Тестът не предоставя отговор дали нещо е интелигентно, а по-скоро се фокусира върху това, дали нещо има културно ориентирана човешка интелигентност.

Тестът на Туринг се използва, за да се провери нивото на логическо мислене на машина, но съзнанието не е само логика. интелигентността не се състои само от логическата и решаващата проблеми част, но също така има и емоционален аспект към нея, например самосъзнание. Във връзка с това Тюринговият тест може да се счита за стъпка в процеса на установяване дали машина е наистина интелигентна, много полезна и важна стъпка, но тя не може и не трябва да бъде единственото условие за класифициране. Също така е важно да се отбележи, че изкуствената интелигентност не е задължително да бъде същата като интелигентността, базирана на човек.

Кибератаки и изкуствен интелект

Изкуствената интелигентност също е внедрена в областта на киберсигурността, където тя става все по-важна за защитата на онлайн системи срещу кибератаки и незаконни опити за достъп. Ако се използва правилно, ИИ системите могат да бъдат обучени да осъществяват автоматично откриване на киберзаплахи, да генерират предупреждения, да разпознават нови видове зловреден софтуер и да защитават важни данни за предприятията.

Използването на методи и алгоритми от изкуствен интелект за защитни цели е в напреднал стадий, но също се отняса и за нападателни приложения. Допълнителното разработване на нападателни технологии с изкуствен интелект може да направи зловредния софтуер по-успешен, като подобри автономността, сложността, скоростта и затрудни откриването му. Бъдещо поколение на зловредни програми може да получи подкрепа от изкуствения интелект и да има способността да функционира независимо. Способността да се адаптира към нова среда, използването на изкуствения интелект за създаване на умни вируси и зловреден софтуер или моделиране на приспособяващи се атаки са допълнителни възможности, които създателите на зловреден софтуер могат да използват. В описание от е представена версия на зловреден софтуер, която е подобрена чрез методите на дълбокото обучение и може да използва разпознаване на лица, глас и местоположение, за да намери и идентифицира своята цел, преди да атакува.

Един интересен аспект на кибератаките е възможното използване на изкуствения интелект за извършване на атаки срещу други изкуствени интелекти. Една от техниките, които може да се използват, се нарича (adversarial inputs) "противникови входове", при които злонамерени лица създават входни данни, които причиняват модели да предсказват неправилно, за да се избегне откриването на истинските отклонения от предварително зададени стойности. В се представя техниката MalGAN, базирана на генеративни противникови мрежи (GAN), която се използва за създаване на противникови примери, които успешно измамват модели за откриване, базирани на тестване тип „черна кутия“, използващи машинно обучение. Това е метод, чрез който злонамерени субекти могат да злоупотребяват с изкуствения интелект, за да извършват нападения срещу други системи с изкуствен интелект.

Изводи и заключения към първа глава

- От направена анализ се установи, че няма единни общо приети определения за понятията киберсигурност и изкуствен интелект. Въпреки това могат да бъдат изведени по – общи дефиниции, чрез които максимално точно и ясно да се представят и предадат концепциите, които се описват с посочените термини;
- Установено е, че с драстичното развитие на информационните технологии се появиха сложни и многопосочни проблеми;
- Компютърната сложност на кибератаките налага разработването на нови стратегии, които са по-надеждни, мащабируеми и приспособими. Както е изследвано, фокусът не следва да бъде само върху технологичния аспект на проблемите, но също и върху моралния и етичния аспект. В противен случай, продължаващото развитие и оръжейна надпревара между изкуствения интелект и киберсигурността може да доведе до свят, в който множество основни свободи се жертват в името на т.нар. повишаване на сигурността.

Втора глава. Анализ на APT хакерските групи. Методи от изкуствен интелект при анализ на зловредни файлове.

Разгледани са алгоритми за машинно обучение, които са приложими при анализ и приписване (attribution to) на зловредни файлове (malware), използвани от Advance

Persistence Threat (APT) хакерски групи, за осъществяване на пробиви в киберсигурността. Процедурата, която ще бъде следвана е следната:

- Първоначално ще се избере множество с данни, върху което да бъдат проведени обучението и експериментите;
- Ще бъдат избрани метрики за оценка;
- За целите на класификацията ще бъдат използвани хеш стойности, генерирани от доказани и общоприети алгоритми за хеширане. В последствие хешираните стойности ще бъдат използвани в новата методология за анализ и приписване на зловредни файлове към хакерските групи и държави-източник;
- Данните ще бъдат разделени на обучаващи (трениращи) и проверяващи (тестови);
- Ще бъдат избрани и разгледани различни видове алгоритми за машинно обучение, в по-голямата си част на база на проучената в първа глава научна литература;

С цел по-голяма яснота при работата с избраните данни и по-ясното им дефиниране и класифициране в необходимите категории е извършен анализ на различни APT хакерски групи.

Същност на APT хакерските групи

APT (Advanced Persistence Threat) атаките представляват сложни и продължителни киберзаплахи, обикновено извършвани от държави или групи, подкрепяни от държавата, които проникват в компютърни мрежи и остават незабелязани за дълго време. Напоследък терминът може също така да обхваща недържавни групи, извършващи целенасочени атаки с определени цели.

Мотивациите зад тези атаки са най-често политически или икономически. Всички основни бизнес сектори са ставали обекти на такива атаки, независимо дали целта е кражба на информация, шпиониране или причиняване на щети. Засегнатите сектори включват правителства, отбранителни компании, финансови и правни услуги, индустрии, телекомуникации и потребителски стоки.

Освен за единични атаки, APT групите може да участват в дълги и координирани киберкампании, често граничещи, ако не напълно съвпадащи, с дефиницията за кибервойна. Съществува значителен дебат сред експертите относно определението за

кибервойна и дори дали такова нещо съществува. Едно мнение е, че терминът е погрешно наименование, тъй като нито една кибератака до момента не може да бъде описана като война. Алтернативно мнение е, че това е подходящ етикет за кибератаки, които причиняват физически щети на хора и обекти в реалния свят.

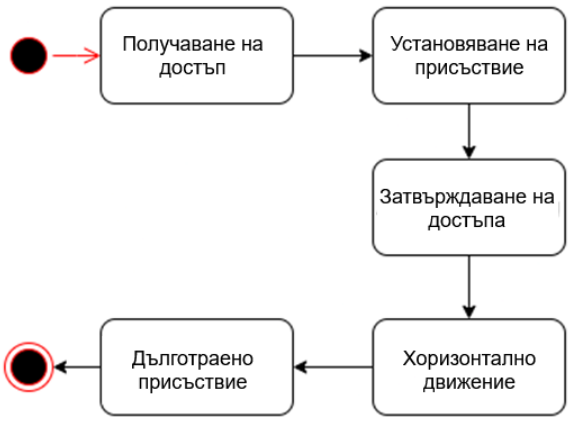
При проучване на публично достъпни бази с информация за APT групи и свързани към тях файлове е установено, че в почти всички от тях попадат групи, свързани с дръжавни структури на Китайската народна република (КНР), Руската Федерация (в частност със специалните служби на страната, напр. ФСБ – Федера́льная слу́жба безопа́сности) и Корейската народно – демократична република (КНДР). Една част от достъпните бази съдържат информация и за APT групи свързани със Съединените американски щати (САЩ) и Пакистан.

Киберсигурност, изкуствен интелект и APT хакерски групи

APT представляват сериозен риск за предприятията по целия свят в динамично развиващата се сфера на киберсигурността. Тези усъвършенствани заплахи експлоатират слабости в системите за продължителен период от време, което ги прави трудни за откриване и неутрализиране. Изкуственият интелект (ИИ) се е утвърдил като ключов инструмент за подобряване на техниките за идентифициране и противодействие на APT атаките, но също така се използва и за разработване на злонамерен софтуер и методи от различни киберпрестъпни групи.

Както е посочено в , APT представляват прикрит злонамерен субект, обикновено държава или спонсорирана от държава организация, която получава неоторизиран достъп до компютърна мрежа и остава скрита за дълъг период. Пример за известна APT група може да бъде даден с посочената в предходната секция група APT29, която се свързва с Руската служба за външно разузнаване (СВР). Групата е активна поне от 2008 г., като често атакува правителствени мрежи в Европа и страните членки на НАТО, както и изследователски институти и аналитични центрове.

Обикновено APT атаките преминават през пет основни етапа при извършване на операции срещу даден противник, като фигура 8 представя графично тези стъпки.



Фигура 8. 5 основни етапа на APT кибероперации

APT групите могат да представляват съществена заплаха към нормалното функциониране на голям обхват комуникационни и информационни системи. Поради тази причина основният фокус на изследването е да се създаде приложение, базирано на методи от ИИ, което да подпомага откриването и атрибуцията на злонамерен APT софтуер.

Избор на множество от данни за обучение

Във всеки проект или процес на вземане на решение началната фаза на събиране на данни може да се счита за една от най-важните, ако не и най-важната. За целта са разгледани множество публично достъпни бази с информация за APT групи, като една от тях е MITRE ATT&CK, която съдържа информация за тактики и техники на проникване и експлоатация на уязвимости, като освен тях може да бъдат намерена и сравнително детайлна информация за хакерски групи и потенциалната им връзка с определени държави. Обикновено информацията е структурирана чрез използването на полетата ID (уникален идентификатор), Name (наименование), Associated Groups (свързани групи) и Description (описание). Информацията се обновява периодично и може да бъде използвана за инициален или част от задълбочен анализ на настъпила кибератака.

Разгледан е и друг набор от данни, който в последствие е използван за нуждите на изследването. Той се състои от приблизително около 3500 маркирани образци на зловреден софтуер, които са свързани с 12 APT групи, за които се твърди, че са спонсорирани от 5 различни национални организации (държави). Те са събрани от

различни източници на информация за заплахи, като търговски бази данни, хранилища с отворен код и вътрешни регистрационни файлове за киберсигурност. Данните са събирани от множество доклади за киберзаплахи и е получен списък на всички хеш стойности на файловете, които в последствие са използвани като индикатори за компрометиране (IoC – Indicator of Compromise). Посочените образци са достъпни в публично git хранилище, от където могат да бъдат изтеглени са последваща обработка.

Данните са представени под формата на .csv файл, в който всеки нов ред представлява нов запис в множеството от данни. Основните характеристики и етикети, избрани за анализ, включват:

- Семейство (Family): Показва семейството на зловредния софтуер, което обозначава неговия тип или произход. Тази характеристика помага за идентифицирането на общи модели сред сходни видове зловреден софтуер;

- Статус (Status): Показва актуалния статус на зловредния софтуер (активен или неактивен), което може да бъде индикатор за наскоро възникнали заплахи или по-стари атаки;

- Криптографски хешове (Cryptographic Hashes - MD5, SHA1, SHA256): Тези характеристики предоставят уникални идентификатори за всяка проба зловреден софтуер. Макар че не показват пряко неговия произход, определени модели в хеш стойностите могат да отразяват специфични инструменти за компилация или библиотеки, използвани от различни заплахи;

- Държава (Country): Този етикет обозначава предполагаемата държава на произход на зловредния софтуер, базирано на разузнавателни доклади и предишни атрибуции. Тази характеристика се използва като основна променлива за задачата по класификация на държавите;

- APT група (APT Group): Този етикет идентифицира конкретния източник на заплаха, в случая спонсорирана от държавата хакерска група, свързана със зловредния софтуер. Тази характеристика се използва като основна променлива за задачата по класификация на APT групите.

Избраните характеристики и етикети са предназначени да създадат пълен профил на всяка проба от зловреден софтуер, включително контекстуални данни (като семейство и статус) и статични атрибути (като криптографски хешове).

Обобщен профил на използваните данни е представен в Таблица 1.

Таблица 1. Сравнителна таблица на част от характеристиките на множеството от данни, съотносими към разглеждания проблем

Държава	Група	Семейство	Брой записи
КНР	APT 1		405
КНР	APT 10	i.a. PlugX	244
КНР	APT 19	Derusbi	32
КНР	APT 21	TravNet	106
Руска Федерация	APT 28	*Bears*	214
Руска Федерация	APT 29	*Dukes*	281
КНР	APT 30		164
КНДР	DarkHotel	DarkHotel	273
Руска Федерация	Energetic Bear	Havex	132
САЩ	Equation Group	Fannyworm	395
Пакистан	Gorgon Group	Different RATs	961
КНР	Winnti		387
Общо			**3594**

Предварителна обработка на трениращото множество

За да се гарантира, че предложеният модел може да се обобщи и прилага за различни заплахи, наборът от данни е внимателно подбран, така че да включва широк спектър от проби, свързани с различни държави и APT организации. Включени са проби от добре познати APT организации, като: APT 28 (Русия), APT 10 (Китай), Equation Group (САЩ).

Целта е да се подобри предсказателката мощ на модела, като той разпознава уникални модели, характерни за различните заплахи. Неравномерното разпределение на класовете е един от основните проблеми при наборите от данни за киберсигурност. Поради голямата активност на някои държави (като Китай и Русия) и APT групи (като APT 28 и APT 10), те са свръхпредставени, докато други заплахи са недостатъчно представени. За минимизиране на този дисбаланс е използвана комбинация от

повторен избор на проби (resampling) и претегляне на класовете (class weighting), като тази тема ще бъде разгледана по-подробно в следващите раздели. За обучение на модела е използвана по-голямата част от данните (около 95%), а за тестване е избрана малка случайна извадка, съставена от неизвестни за модела данни.

Избор на хиперпараметри за машинно обучение

Основна цел на проучването е успешно да се определят, чрез използването на методи от машинното обучение, APT групата и държавата към който принадлежи даден зловреден код, разглеждан в аспекта на киберсигурността. Използваните при експериментите методи от сферата на изкуствения интелект са избрани както от проучените в литературата в първа глава, така и някои допълнителни методи. Основният алгоритъм за класификация с машинно обучение, който ще се ползва по време на експериментите и който ще бъдат проучван в настоящето изследване е Random Forest, развит на базата на Decision Trees.

Самите хиперпараметри в машинното обучение представляват настройваеми параметри, които потребителят може да регулира, за да оптимизира работата на алгоритъма спрямо конкретна задача. За разлика от тях, вътрешните параметри на алгоритъма се определят автоматично по време на обучението. Важно е да се отбележи, че стойностите на хиперпараметрите принадлежат на крайно множество, а не на непрекъснат диапазон. Това означава, че ако даден параметър показва добра генерализация при определена стойност, не е задължително междинните стойности също да осигурят добри резултати. Всяка стойност може да се провери индивидуално в рамките на предварително дефинирания набор от стойности в зависимост от избрания алгоритъм и неговите имплементирани специфики. В проведените експерименти е използван методът Random Search за произволен избор на стойности на хиперпараметри при валидацията от предварително зададено множество.

Алгоритъма Random Forest (произволна гора от дървета на решения) е подходящ за класификационни задачи и се подготвя чрез контролирано (надзиравано) обучение. Накратко този алгоритъм може да бъде обобщен като мета-оценител, който комбинира множество класификатори, базирани на дървета на решенията, приложени върху различни подизвадки от наличните данни. Чрез усредняване на резултатите този метод повишава точността на прогнозиране и намалява риска от прекаленото

наслагване (overfitting), като предотвратява „наизустяването“ на данните и подпомага реалното им обучение.

С цел възможно най – голямо опростяване на решението и най – бързото и лесно имплементиране е взето решение за хиперпараметрите да бъдат използвани стойностите, които са зададени по подразбиране, като в таблица 2 са обобщени избраните хиперпараметри за класификатора Random Forest.

Таблица 2. Обобщени хиперпараметри за класификационния алгоритъм Random Forest

Хиперпараметър	**Стойност**
max_depth	None
n_estimators	100
min_samples_leaf	1
max_features	sqrt(n_features)
criterion	gini
class_weight	Default

Изводи и заключения към втора глава

- Данните за хакерски групи, които са подпомагани и/или свързани с дадени държави, както и извършените от тях кибератаки са разнородни и понякога многобройни за кратък интервал от време;
- Установено е, че наличието на подходящо множеството от данни е ключов фактор за успешното прилагане на алгоритми за машинно обучение при решаване на поставената задача, на база на което е избран оптимален набор от данни, съдържащ данни с информация за връзките между зловредни файлове, семейство зловреден код, хакерски групи и държави;
- Извършена е предварителна обработка на избраното множество от данни, като са подбрани само необходимите за постигането на поставените цели класификатори във формат, подходящ за машинно обучение;
- Избрани са алгоритъм от сферата на машинното обучение и хиперпараметри, чрез които автоматично класифициране на различните категории, може да се извърши на базата на експериментална оценка с трениране и валидация;
- След анализ е взето решение хиперпараметрите да бъдат оставени със стойностите, които са им зададени по подразбиране с цел да се улесни и опрости

имплементацията на решението, а оценката на постигнатите резултати ще бъде извършена основно чрез използването на метода "черна кутия", който се фокусира върху изследване на функционалността на даден програмен продукт.

Трета глава. Модел за машинно обучение за класификация на APT зловредни файлове, използвани в кибератаки.

Класификационните характеристики играят ключова роля при обучението на алгоритми за машинно обучение. Ефективността на алгоритъма зависи от правилния избор на тези характеристики и техните стойности. Поради тази причина, след като суровите данни бъдат подготвени за машинна обработка, е необходимо да се подберат само онези, които ще допринесат за оптималното функциониране на класификационните алгоритми. Това включва:

- Нормализиране и обработка на данните, така че те да бъдат в подходящ формат за обучение, като същевременно се елиминира ненужната или подвеждаща информация;
- Оценяване и преоценяване на различните класове с цел постигане на баланс между тези с по – малък брой записи и тези с по – голямо общо представяне в данните;
- Определяне на най – съотносимите класификационни характеристики.

Оценяване и нормализиране на обучителното множество от данни

Преди да се определят класификационните характеристики, които ще бъдат използвани за обучение на избрания метод от сферата на изкуствения интелект, е необходимо да се намери решение на два проблема, произлизащи от характеристиките на трениращото множество от избраните данни. Първият проблем е свързан с липсата на информация за принадлежност към семейство на зловреден програмен код в 956 от всичките 3594 записа. Вторият проблем е свързан с факта, че определени държави и/или хакерски групи имат неравномерно представяне, като някои се срещат значително по – често, отколкото други.

Данните, които са избрани за анализ в изследването, съдържат следните полета с информация: ID, Country, APT-group, Family, Status, MD5, SHA1, SHA256, Source. При първоначалния преглед на избора на обучителното множество от данни е

установено, че за част от записите липсва информация за принадлежност към определено семейство на зловреден програмен код. След разглеждане на произволен случаен набор от записи е открито, че освен информация за принадлежност към семейство част от множеството съдържа и други непълни данни, като например стойности на хеш функция за алгоритми SHA1 и/или SHA256, което е видно и от фигура 9, която изобразява пет случайно подбрани записа от избраното множество за обучение.

ID,Country,APT-group,Family,Status,MD5,SHA1,SHA256,Source

4,China,APT 1, ,X,00f24328b282b28bc39960d55603e380, , ,https://www.fireeye.com/content/dam/fireeye-www/services/pdfs/mandiant-apt1-report.pdf

1762,Russia,APT 29,CosmicDuke,V,cd012e8f5340d2e148d2c2cbac4270a1,6b7a4ccd5a411c03e3f1e86f86b273965991eb85,
92172ff7bfeee332409a145bc626bebf732225d006877168f35c046368e5118c,https://www.f-secure.com/documents/996508/1030745/dukes_whitepaper.pdf

2193,North-Korea,Dark Hotel,Dark Hotel,V,28b1569109fcae8cfcdcfbe9c4431b66,964c32a7223c8790b05ae94e59c8ca26a78eab67,
276b17244408e7e698e837a0a105c7c3857acfac37e2e837d4b10e6904fd9dc3,https://media.kasperskycontenthub.com/wp-content/uploads/sites/43/2018/03/08070901/darkhotelappendixindicators_kl.pdf

2917,USA,Equation Group,FannyWorm,V,e10a9df3745684581ea3cf5ab22e3e90,d15f43f14372862111560cdcf48fd3c15418f1ea,
5d411570fc456403797add124bcef890238c2c94c7997972db30121132e81a95,https://cantagiodump.blogspot.com/2015/02/equation-samples-from-kaspersky-report.html

4391,China,Winnti, ,X,0996b71f1364acde317881810c5912f0, , ,https://media.kasperskycontenthub.com/wp-content/uploads/sites/43/2018/03/20134508/winnti-more-than-just-a-game-130410.pdf

Фигура 9. Случайно подбрани записи от избраното множество от данни

След извършен анализ е установено, че за спряване с проблема е подходящо записите с липсващи данни да не се премахват, а полетата, в които няма информация, се запълват с класификатор “unknown”. Този подход е избран, тъй като в противен случай голяма част от основното множество трябва да бъде премахната, видно от фигура 10. Това е възможно да породи голям проблем при обучението и последващата работа на алгоритъма, а в сферата на киберсигурността бързите и точни прогнози и решения често са жизненоважни за справяне и/или намаляване на щетите и въздействието на кибернападението.

Друг често срещан проблем е наличието на повтарящи се записи в множеството от данни. При първичен анализ на избраното множество от данни е възможно да бъде направено заключение, че голяма част от записите в него се повтарят, особено ако се обърне внимание на полетата, които са обект на изследване и анализ, а именно “Country” и “APT-group”. Необходимо е да се отбележи обаче, че подобно заключение би било погрешно, тъй като всеки запис от данните съдържа стойност на хеш функция на зловреден софтуер, която е уникална за него и не се повтаря никъде в множеството. На база това е прието, че липсва необходимост да бъдат предприемани мерки за справяне с описания казус, тъй като в настоящия случай той липсва.

Фигура 10. Отношение между записи, описано чрез полето "Семейство"

С цел нормализация на данните и допълнителна оптимизация е взето решение избраното множеството от данни да се раздели на две отделни множества, като от тях отпаднат полетата, които не са жизнено необходими за реализирането на разработваното решение. Първото множество се състои от две полета – “Country” и “APT-group”, извадка от което е представено в таблица 4 и е условно наречено „етикети“ (labels).

Таблица 4. Извадка на част от полетата и записите от множеството „етикети“

Country	**APT-group**
China	APT 10
China	APT 10
China	APT 10
Russia	APT 28
Russia	APT 28
Russia	APT 28
Russia	APT 29
China	APT 30
China	APT 30
China	APT 30
North-Korea	Dark Hotel
North-Korea	Dark Hotel
USA	Equation Group
USA	Equation Group
Pakistan	Gorgon Group

Pakistan	Gorgon Group
Pakistan	Gorgon Group

Второто множество се състои от пет полета – “Family”, “Status”, “MD5”, “SHA1” и “SHA256”, извадка от което е представено в таблица 5 и е условно наречено „характеристики“ (features).

Таблица 5. Извадка на част от полетата и записите от множеството „характеристики“

Family	**Status**	**MD5**	**SHA1**	**SHA256**
o.a. PlugX	V	472b171…	2c1b42e…	472b171…
o.a. PlugX	V	4840ee7…	e113eaf…	4a1c9b9…
o.a. PlugX	X	486a97e…	unknown	unknown
unknown	V	9eebfeb…	e2450df…	e6d09ce…
PinchDuke	V	c166d00…	07b4e44…	b06285f…
PinchDuke	V	2e30fd3…	0cf68d7…	52ba22d…
CosmicDuke	V	351c913…	580eca9…	1590bdb…
CosmicDuke	V	f226063…	5a199a7…	fe5bc12…
CloudDuke	V	f338e21…	78fbdfa…	dea20c2…
CloudDuke	V	e268e5c…	e33e634…	bc20725…
unknown	V	002e279…	b836d5d…	cd2d206…
unknown	V	010ca5e…	aba8b9f…	0beb385…
unknown	V	ff00682…	53ed9b2…	f546f34…
Dark Hotel	V	08b04d6…	42973e5…	da7f9ba….
Dark Hotel	X	08e0852…	unknown	unknown
Dark Hotel	X	091e436…	unknown	unknown
Havex	V	1d6b11f…	7f24973…	0b74282…
Havex	V	68a5f81…	d9bac9d…	ce99e5f…
FannyWorm	V	002f5e4…	90dad7e…	54a0c42…
FannyWorm	V	00f5f27…	50f8342…	ce03f48…
FannyWorm	V	0a78f4f…	621299e…	ead0138…
Different RATs	V	48159a3…	e1391e1…	5cb8b4b…
Different RATs	V	d4addc7…	f0d601d…	ba5afe1…
Different RATs	X	unknown	unknown	d338478…
Different RATs	X	unknown	unknown	8b7ccef…
Different RATs	X	unknown	unknown	4f7049d…

В описаното множество някои държави (напр. КНР и Русия) и APT групи (напр. APT 10, APT 28) са свръх представени поради своята активност, докато други са слабо представени. За да се преодолее този проблем, е използвана комбинация от техники за повторно балансиране на извадките и претегляне на класовете:

- **Повторно балансиране на извадките** – Данните са преработени чрез комбиниране на свръх представяне на извадките за малцинствените класове и намаляване на коефициента на представяне на извадките за свръх представените класове. Този подход е насочен към създаване на балансиран обучителен набор, чрез който се цели предотвратяване на пристрастността на модела към най-често срещаните класове;

- **Претегляне на класовете** – Освен повторното балансиране на извадките, при обучението на модела са приложени и тегла за класовете, чрез които да се постигне високо ниво на увереност, че класификаторът придава подходяща значимост на слабо представените класове. Тази промяна подобрява способността на модела да разпознава по-рядко срещани, но значими файлове със зловреден софтуер.

Избор на алгоритъм от сферата на машинното обучение

Основният алгоритъм, избран за това изследване, е Random Forest – техника за машинно обучение, известна със своята устойчивост и приспособимост при решаване на задачи за класификация, особено в областта на изследването на зловреден софтуер и киберсигурността. Random Forest е избран поради способността му да обработва сложни и многомерни набори от данни, да контролира взаимодействията между характеристиките и да осигурява вероятностни предсказания – всички тези качества са критични при вземането на надеждни решения за атрибуция на зловреден софтуер.

Системна архитектура на експериментален прототип

Валидацията на предложения метод е извършена върху представителна извадка от множеството от данни избрана на случаен принцип. Резултатите са оценени както статистически, така и експертно. За нуждите на процеса на валидация е реализиран експериментален прототип под формата на приложение за класификация на зловреден софтуер, което е проектирано да оптимизира процеса на анализиране на проби от зловреден софтуер, извличане на съответните характеристики и прогнозиране на

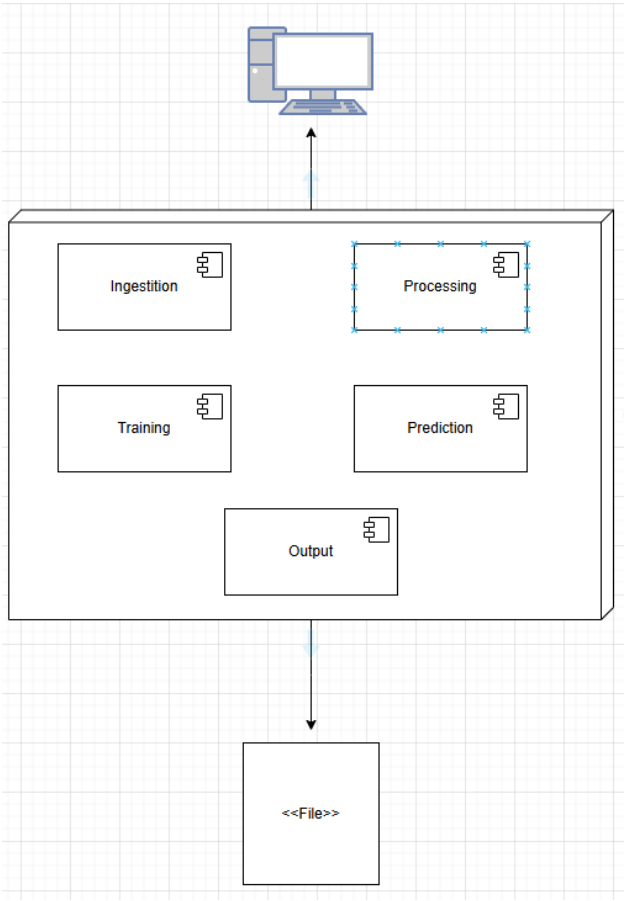
вероятната държава на произход, както и на свързаната група за напреднали постоянни заплахи (APT).

Чрез използването на модулен дизайн, системата комбинира няколко основни процеса, включително: генериране на изходни резултати; класификация чрез машинно обучение; извличане на характеристики; подготовка на данни.

Тази модулна архитектура гарантира гъвкавост, мащабируемост и лесна поддръжка на системата, като същевременно улеснява безпроблемния работен процес, визуализиран на фигура 13, от зареждането на данни до генерирането на предсказания.

Фигура 13. Диаграма тип „System flow“

Предложеното софтуерно решение е реализирано по начин, който да предоставя независимост от гледна точка на интернет свързаността. Всички компоненти и ресурси, които са необходими за неговото функциониране работят върху потребителската машина, както е видно и от диаграмата на фигура 14. Прототипът е реализиран на програмния език Java, с цел безпроблемното осигуряване на платформена независимост.



Фигура 14. Диаграма на разгръщане

Изводи и заключения към трета глава

- Върху множеството от данни са извършени процеси по подготовка, нормализация и оптимизация, които да подпомогнат процесът на обучение на базата на предварително извършен анализ и оценка на спецификата на данните;

- Извършено е подобрение на класификационните характеристики по начин, по който да отразяват по – точно класификационния статут на взаимовръзка между зловреден програмен код и произхождащи хакерска група и спонсорираща държава;

- На база на проведените експерименти е избран класификационен алгоритъм и метод на обработка, които образуват основата на предложеното решение;

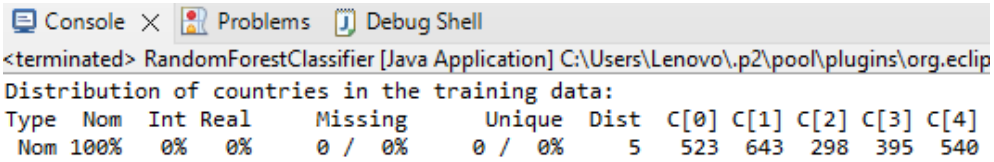
- Създаден е експериментален прототип, което свързва данни, описващи зловреден програмен код, към елементи на класификационната таксономия,

намирайки пълен набор от класове, към които произхожда и/или принадлежат определени хакерски групи и спонсориращи държави;

Четвърта глава. Анализ на предложения модел, резултати и приложение в практиката

Резултати и оценка

Фигура 16 показва първоначалното разпределение на държавите в набора от данни, а фигура 17 – преоценената разпределение на държавите.



Фигура 16. Първоначалното разпределение на държавите.

```
New distribution of countries in the resampled training data:
Type  Nom  Int Real     Missing      Unique  Dist  C[0] C[1] C[2] C[3] C[4]
 Nom 100%   0%   0%     0 /  0%     0 /  0%     5   479  479  479  479  479
```

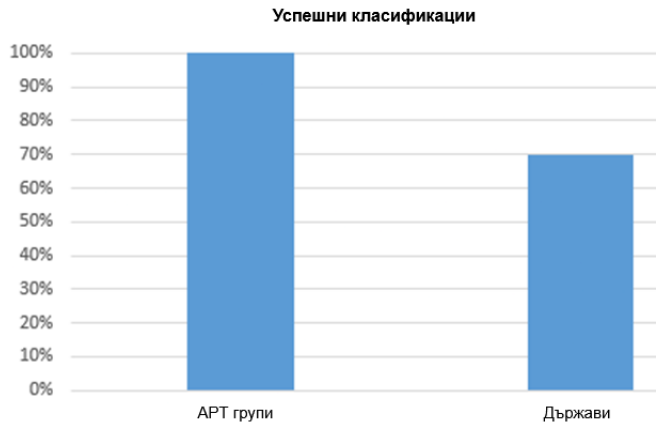
Фигура 17. Преоценената разпределение на държавите

След балансиране на представянето на данните е извършен експеримент за оценка на нивото на ефективност на предложеното решение. Експериментът е проведен по метода на черната кутия, като при първоначалните опити е използван набор от 5 проби на зловреден софтуер, които не са били част от обучителните данни. Данните бяха целенасочено подбрани, така че да представляват различни държави и APT групи. Резултатите от проведен опит са показани на фигура 18.

```
246f88e,predicted country: China, predicted APT-group: APT 10
b8ebff1, predicted country: North-Korea, predicted APT-group: Dark Hotel
:8, predicted country: China, predicted APT-group: Energetic Bear
e946, predicted country: China, predicted APT-group: APT 28
35fe9db610e, predicted country: Pakistan, predicted APT-group: Gorgon Group
```

Фигура 18. Предсказания за APT групи и държави от софтуерното приложение.

Към предишните 5 проби са добавени още 5 на случаен принцип. Резултатите потвърдиха междинния анализ, като точността на APT класификацията достигна 100% или почти 100%, докато точността на атрибуцията по държава е около 70%, както е показано на фигура 19. Причината за намалената точност на атрибуцията по държава се дължи на грешното свързване на определени APT групи, което най-вероятно е резултат от все още не напълно балансирания набор от данни.



Фигура 19. Графично представяне на предсказания за APT групи и държави от софтуерното приложение

Потвърждение на описанието и анализа на получените резултати е видно и от стойностите на метриките Precision, Recall и F1-score за всеки един от класификаторите, представени съответно в таблица 7 и 8.

Таблица 7. Оценка на резултатите за класификатор „Държави“

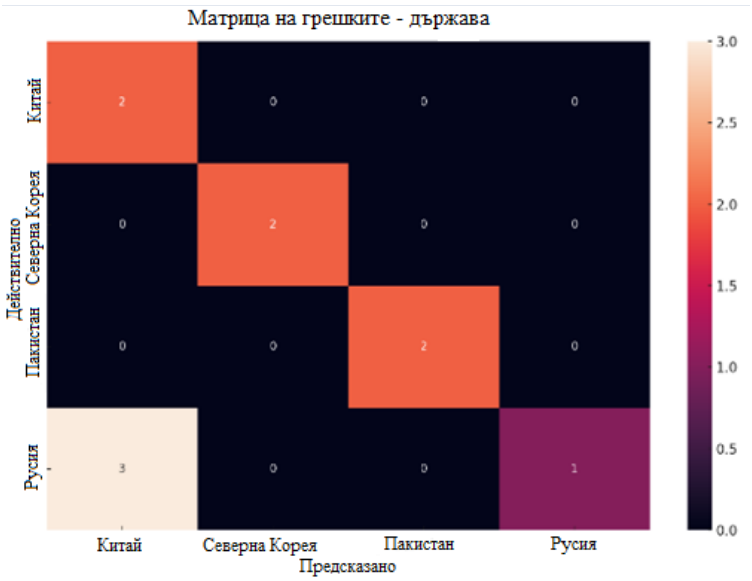
Държава	**Precision**	**Recall**	**F1-Score**
КНР	0.50	0.56	0.53
Руска Федерация	0.50	0.30	0.38
САЩ	1.00	0.33	0.50
Пакистан	0.40	0.50	0.44
КНДР	0.50	0.50	0.50

Таблица 8. Оценка на резултатите за класификатор „APT групи“

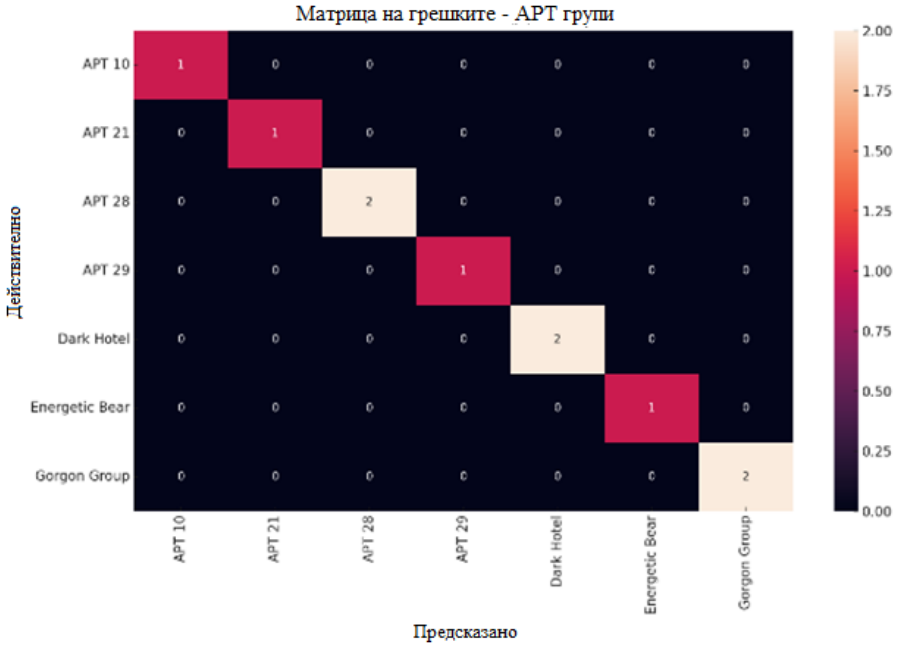
APT група	**Precision**	**Recall**	**F1-Score**
APT 28	1.00	0.50	0.67
APT 29	1.00	1.00	1.00
APT 30	1.00	1.00	1.00

Dark Hotel	1.00	1.00	1.00
Energetic Bear	0.50	1.00	0.67
Equation Group	1.00	1.00	1.00
Gorgon Group	1.00	1.00	1.00

На фигура 20 и 21 са представени Confusion Matrix съответно за класификатор „Държави“ и „APT групи“.



Фигура 21. Графично представяне на предсказания за APT групи и държави от софтуерното приложение



Фигура 21. Графично представяне на предсказания за APT групи и държави от софтуерното приложение

Предложеният модел възприема различен и до голяма степен слабо изследван подход за APT атрибуция. Вместо да разчита на динамичен анализ, поведенчески графове или дълбоки невронни мрежи, той се фокусира върху статични характеристики, извлечени директно от пробите на зловредния софтуер. Извършването на анализ върху големите набори от зловреден софтуер с разглеждания метод е изключително полезно, тъй като той позволява по-бързо обработване на данните и по-малко изчислителни разходи.

Изводи и заключения към четвърта глава

- Валидацията на предложеното решение показва, че успешно се класифицират APT групите при информация за възникнал киберинцидент спрямо предварително проведено обучение върху класификационна таксономия;

- Валидацията при класификацията на държавите показва по-малък процент на успеваемост, при предварително проведено обучение върху същата класификационна таксономия, както при APT групите, като ясно е изведена потенциалната причина за това и е предложено решение, което да намали и/или ограничи откритото отклонение;

- Открито е, че точността на предсказанията на предложеното решение за класификацията на хакерски групи достига 100% или почти 100%, докато точността на атрибуцията по държава е около 70%. Описаните стойности благоприятстват употребата на предложения метод за обогатяване на информацията при оценка на критичността и извършването на анализ на настъпила кибератака;

- Изборът на класове, към които принадлежи хеш стойност на програма със зловредни програми инструкции, е индивидуален за всеки отделен случай, без необходимост от предварително задаване;

- Експерименталният прототип и реализираните в него модули, осигурят удобен достъп за потребителите без да е необходимо да се инсталира допълнителни софтуерни пакети на техните машини.

Научни, научно-приложни и приложни приноси.

В резултат на проведените изследвания са постигнати следните основни приноси:

Научни приноси

- Предложен е модел базиран на класификатор Random Forest за присвояване на клас на принадлежност на зловреден код към определена държава, спонсориращата дейности и развитие на хакерски групи, насочени срещу комуникационна и информационна инфраструктура на други държави и/или организации;

Научно-приложни приноси

- Предложено е използването на таксономия за извършване на прецизиращи процедури при обработката на данни свързани със зловреден код, хакерски групи и спонсориращи държави, избрана на базата на проучване на международни практики в класифицирането на хакерски групи и определянето на инциденти в киберсигурността;

- Предложено е използването на оценъчни техники за оптимизация на класификационните характеристики и избор на хиперпараметри за класификатора Random Forest, които оценяват свръх представяне на извадките за малцинствените класове и намаляване на коефициента на представяне на извадките за свръх представените класове;

Приложни приноси

- Разработен е експериментален прототип, използващ предложения модел с класификатор Random Forest, който реализира автоматизиран процес по класифициране на характеристики на файлове, съдържащи зловреден код към определение хакерски групи и/или спонсориращи държави и улеснява дейностите за анализ и оценка на експерти по киберсигурност;

- Предложеният модел е експериментално оценен с използване на разработения експериментален прототип за валидиране на възможностите на модела за подобряване на класификацията на файлове и улесняването на работата на експерти по киберсигурност.

Списък с публикации по дисертационния труд

1) Grozdan Hristov, Adelina Aleksieva-Petrova and Ivan Stankov, "CyberAttacks and Artificial Intelligence: A Systematic Mapping," *2023 International Scientific Conference on Computer Science (COMSCI)*, Sozopol, Bulgaria, 2023, pp. 1-7, doi: 10.1109/COMSCI59259.2023.10315929.; (Индексирана в Scopus)

2) Ivan Stankov and Grozdan Hristov, "CYBERSECURITY AND INFORMATION SECURITY", Youth Scientific Conference "Machines Innovation and Technology", 2020,

i. Цитирана в:

1. Suraeva, M. & Afonasyev, M. & Kucheryavenko, D.. (2022). The Impact of Digitalization on Innovative Approaches to Economic Security in Regions. 10.1007/978-3-030-83175-2_43.;

3) Ivan Stankov and Grozdan Hristov, "KNOWLEDGE MANAGEMENT INFORMATION SYSTEMS IN EDUCATION AND SCIENCE", International Scientific Conference Computer Science', 2020 (Индексирана в Scopus)

i. Цитирана в:

1. X. Gong, H. Li, J. Gui, Z. Shu and T. Huang, "Resilient and Efficient Multirobot Pickup and Delivery Against Strategic Attacks: A Three-Layered Framework," in IEEE Internet of Things Journal, vol. 12, no. 20, pp. 42317-42336, 15 Oct.15, 2025, doi: 10.1109/JIOT.2025.3593317.;

2. D. Andreev, R. Trifonov and M. Lazarova, "Challenges Regarding AI Integration in V2X Communication," 2024 12th International Scientific Conference on Computer Science (COMSCI), Sozopol, Bulgaria, 2024, pp. 1-5, doi: 10.1109/COMSCI63166.2024.10778510.

4) Grozdan Hristov, Ivan Stankov and Dayana Mladenova, "CyberAttacks and robotics," *2023 31st National Conference with International Participation (TELECOM)*, Sofia, Bulgaria, 2023, pp. 1-4, doi: 10.1109/TELECOM59629.2023.10409708. (Индексирана в Scopus)

5) Grozdan Hristov. "AI-enhanced cybersecurity: Machine learning classification application for APT malware attribution." International Journal of Innovative Research and Scientific Studies. Vol 8 (1). pp. 2295-2304. 26.02.2025. 10.53894/ijirss.v8i1.4955.

Abstract of PhD thesis

Grozdan Kostadinov Hristov, M.Sc.

Methods and algorithms from machine learning in cybersecurity

The focus of the dissertation work is done on the use of artificial intelligence in the field of cybersecurity. The research is done in regards to the profile of advanced persistent threat (APT) groups and how methods from the field of machine learning can be utilized for faster analysis of malware files, that are or have been used by said groups.

One of the research focuses is on an analysis of the profile of APT groups. This is done in order to desern how to optimize their classification and which characteristics can be used in order to achive the said goal. To get a broader understanding and accelerate the research process different knowledge bases are taken into account and analyzed, including MITRE ATT&CK which has an extensive and indepth taxonomy regarding APT groups as well as their tactics, techniques and procedures. For future classification a data set was chosen, containing approximately 3,500 tagged malware samples, which are associated with 12 APT groups, allegedly sponsored by 5 different national organizations (states). The specified samples are available in a public git repository, from where they can be downloaded for further processing.

The other research focus is on the use of machine learning methods and algorithms for classification of malware files. An analysis is done after which it was decided to use the algorithm Random Forest for the task of classification of APT malware files. An experimental prototype is created which implements a model leveraging static features extracted from the malware, including cryptographic hash values (MD5, SHA1, SHA256) and malware family labels, to build robust Random Forest classifiers for country of origin and APT attribution. The choice of static analysis allows for efficient and scalable feature extraction, making the approach well-suited or large-scale datasets and real-time applications. The experimental results show an achievement for APT accuracy reaching 100% or very close to 100%, while the country accuracy was around 70%.

In conclusion the proposed solution does have drawbacks when it comes to country attribution, but this is negated by the accuracy of the APT group attribution, because this information, in certain cases, is more valuable and can be used for crosschecking different sources to pinpoint the originating country, if there is one.