ТЕХНИЧЕСКИ УНИВЕРСИТЕТ – ВАРНА

Факултет по Изчислителна Техника и Автоматизация

Катедра “Компютърни науки и технологии”

инж. Димитър Георгиев Тодоров

ИЗСЛЕДВАНЕ НА МЕТОДИ ЗА МАШИННОТО ОБУЧЕНИЕ ЗА КРИПТИРАНЕ НА ИНФОРМАЦИЯ

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”

Докторска програма: “Компютърни системи, комплекси и мрежи”

Научен ръководител: **доц. д-р инж. Милена Карова**

Рецензенти:

1.
2.

Варна, 2022 г.

Дисертационният труд е обсъден на 27.04.2022г. в катедра „КНТ“ на катедрен съвет, съгласно заповед на Ректора на ТУ-Варна № ..... /……..... г. и насочен за защита.

Автор: Димитър Георгиев Тодоров

Заглавие: Изследване на методи за машинното обучение за криптиране на информация.

ТЕХНИЧЕСКИ УНИВЕРСИТЕТ – ВАРНА

Факултет по Изчислителна Техника и Автоматизация

Катедра “Компютърни науки и технологии”

инж. Димитър Георгиев Тодоров

ИЗСЛЕДВАНЕ НА МЕТОДИ ЗА МАШИННОТО ОБУЧЕНИЕ ЗА КРИПТИРАНЕ НА ИНФОРМАЦИЯ

А В Т О Р Е Ф Е Р А Т

на дисертация за придобиване на образователна и научна степен

“**ДОКТОР**”

Варна, 2022 г.

Дисертационният труд съдържа 137 страници, включително 64 фигури, 10 графики и 16 таблици, оформени в 4 глави за решаване на формулираните основни задачи, списък на основните приноси, списък на публикациите по дисертацията и използвана литература от 121 заглавия, от които 11 на кирилица (8 на руски език и 3 на български език) и 110 са на латиница (английски език), от които 5 са интернет адреси.

Защитата на дисертационния труд ще се състои на ..... г. от .....ч. в ..... на открито заседание на жури сформирано със заповед на Ректора №...../..... г.

Материалите по защитата (дисертацията, рецензиите и становищата) са на разположение на интересуващите се в Докторантски център, стая 318 НУК.

ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

1. Актуалност на проблема

Сигурността е много важна част от характеристиките на всяка съвременна компютърно-комуникационна система. Известни са различни техники и технологии за нейното осигуряване, като се започне от контрол на достъпа, удостоверяване, оторизация, системи за откриване или предотвратяване на проникване, защитни стени, тунелиране и се стигне до един от основните и способи - криптографията. Тя осигурява средства за автентикация, доказване на интегритет и осигуряване на конфиденциалност. Нейни основни области са симетричното и асиметричното криптиране. Те са средство за защита на информация и комуникации посредством таен ключ. Основна характеристика на всяка една такава система е устойчивостта й на атака по метода на грубата сила – атака с всички възможни варианти на ключа, което определя огромното значението на неговата дължина. С огромния напредък на компютърните технологии през последните години и повишаването на изчислителната им мощност, тази устойчивост намалява. Може да се определи, че тази тенденция определено оказва неблагоприятно влияние върху симетричната криптография, която използва един секретен ключ за конкретен повод. Тя е широко разпространена благодарение на своята бързина и по-лесното си внедряване в различни системи. Това поражда проблем с устойчивостта на симетричните криптографски алгоритми, чието основно решение през годините е увеличаването на дължината на секретния ключ. В съвременните условия трябва да се обърне внимание и на други възможни решения. Едно такова е възможностите на машинното обучение да осигури средство за решаване на проблема - средство, даващо възможност за използването на повече от един секретен ключ.

2. Цел на дисертационния труд, основни задачи и методи за изследване

За постигане на по-висока устойчивост на симетричните криптографски алгоритми **цел** на научните изследвания в дисертационния труд е да се проучат, анализират и изследват методи и алгоритми за машинно обучение за анализ и разпознаване на вида на алгоритъма, към който принадлежи даден секретен ключ, посредством обучаващи данни подходящи по тип, вид и обща големина.

За постигане на поставената цел следва да бъдат изпълнени следните **задачи**:

1. Да се проучат съществуващите методи и алгоритми за машинно обучение за извличане на знания и анализ на масиви от данни.
2. Да се анализират и подберат подходящ по тип, вид и обща големина обучаващи данни за алгоритмите на машинно обучение. Да се предложи алгоритъм, осигуряващ подходящ вид на входните данни за алгоритмите на машинно обучение.
3. Да се проектира и разработи софтуерен модел (приложение) за разпознаване на симетрични криптографски алгоритми чрез машинно обучение, което да се използва за провеждане на експериментални проучвания, както за целите на дисертацията, така и за бъдещи потребности в учебни и приложни процеси.
4. Да се оцени експериментално възможността за използване на методи и алгоритми за машинно обучение за обработване, анализ и класификация на

криптографския ключ и увеличаване на устойчивостта на симетричните криптографски алгоритми.

Научните **методи**, които са използвани, са: проучване на литературни източници в областта на машинното обучение, криптографията и алгоритмите; проектиране на алгоритъм за представяне и предварителна обработка на данните във вид удобен за работата на алгоритмите за машинно обучение; проектиране и имплементация на подход за решаване на проблеми в симетричната криптография, използвайки предложеният алгоритъм за поставяне на данни в еднородна среда и алгоритми за машинно обучение за разпознаване на вида на използвания симетричен криптографски алгоритъм, с който е генериран даден секретен ключ; провеждане на експеримент за оценка, валидация и верификация на предложения подход; статистически методи за обработка и представяне на получените експериментални данни.

3. Обект и предмет на изследването

Обект на научните изследвания в дисертационния труд е устойчивостта на симетричните криптографски алгоритми.

Предмет на научните изследвания в дисертационния труд е повишаването на устойчивостта на симетричните криптографски алгоритми.

Основната изследователска **хипотеза** в дисертационния труд е възможността за повишаване на устойчивостта на симетричните криптографски алгоритми с използване на алгоритми за машинно обучение.

4. Място на изследване

Изследванията и обработката на резултатите са проведени в лабораторната база на катедра КНТ при ТУ – Варна и лабораторната база на докторанта.

5. Научна новост

Предложен е алгоритъм за поставяне на данни в еднородна среда, осигуряващ подходящ вид на входните данни за алгоритмите на машинно обучение. Предложен е подход за формиране на обучаващо множество данни за класифициране на криптографски данни с помощта на предложения алгоритъм за поставяне на данни в еднородна среда. Предложен е подход за проектиране, конфигуриране и имплементиране на модел за подготовка на криптографски данни за работа с алгоритми от машинно обучение. На базата на обстойно проучване са избрани методи и алгоритми за машинно обучение подходящи за имплементиране и реализиране в едно с предложеният алгоритъм за поставяне на данни в еднородна среда в система за разпознаване на криптографски данни.

6. Практическа приложимост

Предложеният подход за решаване на проблеми в симетричната криптография чрез машинно обучение и предложеният алгоритъм за поставяне на данни в еднородна среда дава висока точност при класифицирането на криптографските данни и може да се използва в реален продукт, достъпен за потребителско използване за обмен на криптирана информация и криптиране с различни алгоритми в единна среда. На базата на експериментални изследвания са определени подходяща конфигурация и параметри на модел за класификация на криптографски данни с цел увеличаване на устойчивостта на

симетричните криптографски алгоритми. Предоставя се средство за реализиране на модел на многопрофилно криптиране или криптиране с различни алгоритми в единна среда.

7. Апробация

Основните етапи от разработването на теоретични и приложни резултати на дисертационния труд са докладвани и публикувани в следните научни форуми и издания:

Конференции:

- 1 доклад на международна научна конференция "International scientific-practical conference of young scientists, graduate students and students", Харков, Украйна, 09-10 Юли 2017;
- 1 доклад на международна научна конференция „Advances in Neural Networks and Applications'2018", Св.Константин и Елена, България, 15-17 Септември 2018, **Scopus**;
- 1 доклад на международна научна конференция „International Conference Automatics and Informatics'2021", Варна, България, 30 Септември - 2 Октомври 2021, **Scopus**;

Списания:

- 1 статия в Списание „Computer Science and Technologies Journal", Година XVIII, Брой 1/2020, ТУ-Варна;
- 2 статии в Годишник на ТУ-Варна „Annual Journal of Technical University of Varna";

8. Публикации

Основни постижения и резултати от дисертационния труд са публикувани в 6 научни статии, като 1 от тях е самостоятелна. Научните статии са представени и публикувани в национални и международни реферирани и индексирани издания. Списък на публикациите е приложен в края на автореферата.

СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

Глава 1. Машинно обучение и симетрична криптография. Област на взаимодействие. Организация и изпълнение на съвместни задачи.

Първа глава на дисертационния труд разглежда устойчивостта на симетричните криптографски алгоритми, като обект и анализ на машинното обучение, като средство за повишаване на устойчивостта на симетричните криптографски алгоритми. Направен е обзор на различни литературни източници и е обоснована актуалността на проблема за увеличаване на устойчивостта на симетричните криптографски алгоритми. Главата е организирана в четири основни части. В първа част се разглеждат устойчивостта на симетричните криптографски алгоритми и хибридните криптографските алгоритми, като средство за решаване на проблеми в симетричната криптография. Втора част е свързана с методите и алгоритмите на машинното обучение. Третата част е фокусирана върху практическото приложение на машинното обучение. В четвъртата част се разглежда приложението на машинното обучение в криптографията.

Разпознаване и възстановяване на симетрични секретни ключове чрез електромагнитен анализ и подход с машинно обучение.

В публикации „On features suitable for power analysis - Filtering the contributing features for symmetric key recovery“ [85] и „Machine-Learning-Based Side-Channel Evaluation of Elliptic-Curve Cryptographic FPGA Processor“ [84] е предложено решение за разпознаване и „възстановяване“ на симетричен секретен ключ. Разпознаването е реализирано посредством електромагнитен анализ на хардуерна реализация на криптографски алгоритми AES, DES и RSA и извличане на характеристики с помощта на подход с машинно обучение. За класификация на извлечените характеристики са използвани алгоритми от машинно обучение, като Random Forest, Support Vector Machine и Naive Bayes.

Декларирани са следните резултати на успешно разпознаване съответно на 2, 3 и 4 бит в проценти и изразходваното време:

- Support Vector Machine – 2 бит – 45,7% и време 0,35 с., 3 бит – 49,3% и време 0,34 с., 4 бит – 55,4% и време 0,3 с.;
- Random Forest - 2 бит – 58,0% и време 0,63 с., 3 бит – 56,9% и време 0,64 с., 4 бит – 79,2% и време 0,74 с.;
- Naive Bayes - 2 бит – 55,7% и време 0,06 с., 3 бит – 57,0% и време 0,07 с., 4 бит – 52,5% и време 0,08 с.;

Идентифициране на симетрични алгоритми чрез прилагане на CNN към следи, извлечени от IPT.

През 2021 г. У. Янг и Я. Парк представят нов метод, чрез който да се идентифицират симетрични криптографски алгоритми чрез прилагане на CNN (convolution neural networks - конволюционна невронна мрежа) към следите, извлечени от IPT (Intel Processor Trace) [118]. Конволюционните невронни мрежи са алгоритми за дълбоко обучение с по-голяма производителност в сравнение със съществуващите алгоритми за машинно обучение при класификацията на изображения [118]. Intel Processor Trace е разширение на архитектурата на процесорите Intel, която позволява да се извлича информация за изпълнението на софтуера и има предимството да извлича точна следа на дадена програма чрез заобикаляне

на техниката за отстраняване на грешки. Разгледано в общ план следата, криптирана от алгоритмите със симетричен ключ, се извлича с помощта на IPT. След това се преобразува в изображение, което да бъде вход в конволюционната невронна мрежа. Проведени са експерименти с два различни набора от данни. Първият набор от данни съдържа следи, извлечени от различни типове алгоритми със симетрични ключове – AES, BF, CAST, DES, DES3, IDEA, RC2, RC4, SEED [118]. Резултатите от обучението са класифицирани в девет класа със 100% точност. Вторият набор от данни съдържа следи, извлечени от различните симетрични алгоритми с различна дължина на секретния ключ. Резултатите от обучението са класифицирани в 36 класа с точност 70,55%.

Изводи към първа глава

В резултат на направените проучвания в областта на използване на машинното обучение в криптографията могат да се направят следните по-важни изводи:

1. Симетричните криптографски алгоритми са по-бързи от асиметричните, което е тяхното основно предимство при използването им в криптографията.
2. Основно предизвикателство при използването на симетричните криптографски алгоритми е тяхната устойчивост.
3. С напредването на компютърните технологии съпроводено с няколкократното увеличение на тяхната изчислителна мощност през последните две десетилетия устойчивостта на криптографските алгоритми на атаки по метода на грубата сила намалява.
4. За подобряване на устойчивостта на симетричните крипотграфски алгоритми се използват различни средства, включително и алгоритми за машинно обучение.
5. Различни алгоритми за машинно обучение могат да бъдат използвани за решаване на следните проблеми в симетричната криптография в посока увеличаване на устойчивостта им:

5.1. Проблемът с необходимостта от размяната на секретен ключ, което води до необходимостта от сигурен канал за предаването му.

5.2. Проблемът с необходимостта от секретен ключ за всеки отделен комуникационен канал.

5.3. Проблемът с използването на все по-големи по битова дължина секретни ключове за симетричните криптиращи алгоритми.

6. При повишаване на устойчивостта на симетричните криптографски алгоритми, контролираните алгоритми от машинното обучение са значително по зависими от човешка намеса поради необходимостта за осигуряване на обучителни данни. Но от друга страна с тях може да се осигури по-добро бързодействие при използването им с подходящи по тип данни, тъй като изискват по-малко на брой итерации при изпълнението си и са по-лесни за разбиране и прилагане.

На базата на направените проучвания в областта на синхронната криптография и машинното обучение, както и посочените изводи може да бъде формулирана следната **цел на научните изследвания в дисертационния труд**:

• За постигане на по-висока устойчивост на симетричните криптографски алгоритми да се проучат, анализират и изследват методи и алгоритми за машинно обучение за анализ

и разпознаване на вида на алгоритъма, към който принадлежи даден секретен ключ, посредством обучаващи данни подходящи по тип, вид и обща големина.

За постигане на поставената цел следва да бъдат изпълнени следните задачи:

1. Да се проучат съществуващите методи и алгоритми за машинно обучение за извличане на знания и анализ на масиви от данни.
2. Да се анализират и подберат подходящи по тип, вид и обща големина обучаващи данни за алгоритмите на машинно обучение. Да се предложи алгоритъм, осигуряващ подходящ вид на входните данни за алгоритмите на машинно обучение
3. Да се проектира и разработи софтуерен модел (приложение) за разпознаване на симетрични криптографски алгоритми чрез машинно обучение, което да се използва за провеждане на експериментални проучвания, както за целите на дисертацията, така и за бъдещи потребности в учебни и приложни процеси.
4. Да се оцени експериментално възможността за използване на методи и алгоритми за машинно обучение за обработване, анализ и класификация на криптографския ключ и увеличаване на устойчивостта на симетричните криптографски алгоритми.

Глава 2. Подход за решаване на проблеми в симетричната криптография чрез машинно обучение и алгоритъм за поставяне на данни в еднородна среда.

Втората глава на дисертацията представя разработен подход за реализиране на криптографски задачи чрез машинно обучение. Описва начин за използване на предимствата на симетричната криптография и увеличаване на устойчивостта и на атаки по метода на грубата сила. Организирана е в три основни части. Първата част е посветена на проучване и избор на подходящи алгоритми за машинно обучение, представени са избраните алгоритми от машинно обучение за постигане на поставената цел и е описан начина им на работа. Втората част разглежда избора на симетрични криптографски алгоритми. В нея се описва работата на избраните алгоритми от симетричната криптография. Последната част е фокусирана върху определянето на подходящия вид на данните, с които ще се работи, начина на тяхното съхранение и представянето на алгоритъм за поставяне на данни в еднородна среда (Фиг. 12).

На фиг. 12 е представен общия модел за използване на алгоритми за машинно обучение за решаване на проблеми в симетричната криптография.

Фиг. 12 Взаимодействие между алгоритмите

2.1 Обобщен подход за решаване на основен проблем в симетричната криптография чрез машинно обучение.

Предложеният подход за решаване на проблема с повишаване на устойчивостта на симетричната криптография срещу атаки по метода на грубата сила, се състои в четири основни етапа:

Етап 1: Избор на алгоритъм/и от машинно обучение

Етап 2: Избор на криптографски алгоритъм/и

Етап 3: Предварителна обработка на данните

Етап 4: Класификация на данните

Избора на алгоритмите от машинно обучение и симетрична криптография са основни фактори при определянето на конкретните параметри за предварителната обработка на данните. Така например вида на алгоритмите от машинно обучение обуславят, както типа на данните, така и тяхното съхранение като базови такива. Дължината на секретния ключ използван от дадения симетричен криптографски алгоритъм определя дължината на представените данни. Предварителната обработка на данните е най-важната част от предложения подход. Тя съдържа избора на подходящ вид на данните и алгоритъм за поставяне на тези данни в еднородна среда.

2.2 Избор на алгоритми от машинно обучение

За постигане на основната цел и решаване на задачите на дисертацията е избрано използването на контролирани алгоритми от машинното обучение. Съображенията довели до този избор са следните:

- С помощта на контролирано обучение моделът може да предвиди резултата въз основа на предишен опит, което е в съответствие с обекта и предмета на изследванията в дисертационния труд, изискващи постигане на резултати на базата на предишен опит с базови данни.
- В контролираното обучение може да има точна представа за класовете обекти – данните, които ще се използват са точно и ясно класово определени.
- Контролираното обучение не може да предвиди правилния изход, ако тестовите данни са различни от учебния набор от данни – при използваните данни за разглежданите проблеми с устойчивостта на алгоритмите за симетрична криптография такава разлика няма.
- Алгоритмите за контролирано обучение са по-лесни за реализиране в сравнение с другите алгоритми за машинно обучение, което способства да няма „излишно“ усложняване на системата.
- Преход на реализирана система с контролирано машинно обучение към друг тип машинно обучение в последващи реализации е постижим и лесен са осъществяване.

Контролираното обучение обуславя все пак и външен контрол на една такава система, определящ целенасочеността й. Всички тези доводи определят контролираното обучение като напълно подходящо за вида на решаваните задачи за постигане на целта на дисертацията. Избрани са два от най-широко използваните и най-дискутирани представители на контролираното машинно обучение, които съответстват на изискванията на разглежданите проблеми – алгоритъма на k-а на брой най-близки съседи (на английски език – k-Nearest-Neighbour, или съкратено kNN) и алгоритъма на машина на поддържащите вектори (на английски език – Support Vector Machine, или съкратено SVM).

В Таблица 4 е представено сравнение на няколко различни алгоритъма за контролирано машинно обучение: kNN, SVM, Linear Regression, Naive Bayes, Decision Tree. Сравнението се базира на няколко различни критерия: работа с двоични данни, линейна зависимост между входни и изходни данни, шумоустойчивост, възможност за класификация на текст, работа с малко на брой класове.

Таблица 4: Сравнение между различни алгоритми от контролирано машинно обучение

	работа с двоични данни	линейна зависимост м/у вход - изход	шумоустойчивост	класификация на текст	работа с малко на брой класове
kNN	мн.добра	няма	зависима от k	да	да
SVM	мн.добра	няма	зависима от kernel function	да	да
Linear Regression	добра	голяма	предварителна обработка	не	да
Naive Bayes	мн.добра	малка	зависи от броя (честотата) на атрибутите в набора от данни	да	подходящ за многокласови
Decision Tree	мн.добра	няма	зависи от броя (честотата) на атрибутите в набора от данни	не	да

Представеното сравнение на алгоритмите за контролирано машинно обучение обосновава избора на двата алгоритъма – kNN и SVM, които най-добре отговарят на изискванията и особеностите на вида на данните и обекта на изследване в дисертационния труд.

И двата алгоритъма (kNN и SVM) показват добри резултати при работа с двоични данни. Нямат линейна зависимост на изхода спрямо входните данни, за разлика от Linear Regression. Шумоустойчивостта им е добра и е зависима само от техните основни параметри за работа, за разлика от Naive Bayes и Decision Tree, които са зависими от честотата на атрибутите в набора от данни, а Linear Regression се нуждае от предварителна обработка. Работят добре при класификация на текстови данни и дават много добри резултати при използването на малък брой класове.

2.3 Подбор на симетрични криптографски алгоритми

За целите на дисертацията са избрани да се използват четири широко известни и познати симетрични алгоритъма – AES (Advanced Encryption Standard), DES (Data Encryption Standard), TripleDES (Triple Data Encryption Standard) и RC2. Изборът на тези алгоритми е съобразен със следните изисквания към тях, дефинирани съобразно обекта на изследване и целта на дисертацията:

- Да са познати и широко използвани на практика – и четирите алгоритъм отговарят на това условие.
- Да позволяват използването на различни дължини на секретния ключа – алгоритмите са с различни дължини на ключ: AES позволява при имплементирането си да използва размер от 128 до 256 бита, DES – само 64

бита, TripleDES – от 128 бита до 192 бита на стъпки от 64 бита, RC2 – от 128 до 256 бита.

- Да позволяват интерпретиране в избраната програмна среда за реализация – и четирите алгоритъма са част от платформата .NET Framework и езика C#.
- Да позволяват хардуерна реализация – AES и DES вече имат такава.

В Таблица 5 е представено сравнение на няколко различни симетрични криптографски алгоритми: AES, DES, TripleDES, RC2, IDEA, CAST, Blowfish и ГОСТ 28147-89 според няколко критерия: дължина на ключа, брой тактове за 1 вътрешен цикъл, средно време за генериране на ключ, средно време за криптиране, средно време за декриптиране, наличие на хардуерна реализация и на софтуерна имплементация в .Net Framework. Критериите за сравнение се базират на

Таблица 5: Сравнение на симетрични криптографски алгоритми

	AES	DES	Triple DES	RC2	IDEA	CAST	Blowfish	ГОСТ 28147-89
дължина на ключа (от-до)	128, 192, 256	64	128-192	64-128	128	40-256	до 448	256
брой тактове за 1 вътрешен цикъл в/у Intel Pentium реализация (такта/байт)	18	45	108	-	50	-	18	-
средно време за генериране на ключ (милисекунди)	0,290 0,310 0,340	0,702	0,635	0,488	0,602	0,65	0,512	-
средно време за криптиране	374	389	452	480,7	339,58	208,14	60,3	-
средно време за декриптиране	242	246	275,6	313,2	330,91	264,22	18,72	-
хардуерна реализация	да	да	да	-	да	-	да	-
вкл. в .Net Framework	да	да	да	да	не	не	не	не

изискванията на дефинираната цел на изследванията в дисертационния труд. Четирите избрани и описани алгоритъма са избрани за изследванията в дисертацията поради възможността за реализация и провеждане на изследвания както на различни по дължина ключове, така и на ключове с еднаква дължина. Времето за генериране на ключа за всеки от тях е под 1 милисекунда и е в диапазон от около 0.3 милисекунди. Времето за криптиране и декриптиране е в рамките на около 100 милисекунди, което съпоставено на изчислителните мощности на съвременните компютърни системи може да се приеме за приблизително еднакво, което позволява експерименталните изследвания да са независими от вида на използвания алгоритъм. Алгоритмите имат хардуерна имплементация, с

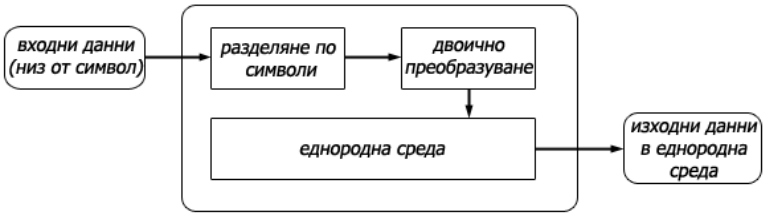
изключение на RC2, за който не е намерена такава информация. И четирите алгоритъма са имплементирани и са част от платформата .Net Framework.

2.4 Алгоритъм за поставяне на данни в еднородна среда

Важен аспект във всяка система с машинно обучение е вида на данните, с които се обучава. Вида на обучаващите данни е в пряка пропорционална зависимост с изходните резултати на такава система. Например една система, която е конструирана за разпознаване на геометрични фигури в изображения, не може да се приложи за разпознаването на текст. Тренировъчните данни определят целенасочеността на системата и задават нейните реални граници на действие. Зависимостта на алгоритмите от машинното обучение от данните е съществена. Не без основание може да се определи, като генетична. Това се обуславя от характеристиките, които притежава даден алгоритъм.

Етапът за предварителната обработка на данните е ключов и най-важен за постигане на целта на дисертацията и решаването на поставените задачи. За предварителна обработка на данните в дисертационния труд се предлага използване на алгоритъм за поставяне на данните в еднородна среда (Фиг. 47), чрез който се постига универсалност на представянето на входните данни по отношение на изискванията на алгоритмите от машинно обучение и еднаквост по отношение на критериите към начина на изразяване на данните.

Алгоритъмът има за цел да преобразува входна последователност от данни, представени чрез символи, които могат да са от различен тип в изходни данни с еднородно представяне.



Фиг. 47 Общ вид на алгоритъма за поставяне на данни в еднородна среда

Подбор на подходящи данни. На базата на изложеното могат да се определят следните няколко критерия за избора на подходящи данни за данните, които се използват за изследванията в дисертационния труд:

- Да използват малък брой символи;
- Да са двупосочно лесни за конвертиране спрямо реалните данни – низ от символи на секретния ключ;
- Да притежават универсалност;
- Да позволяват хардуерна реализация;
- Да позволяват съхраняването им във файлове с прост формат;
- Да позволяват съхраняването им във файлове с малък размер;
- Да позволяват съхраняването им във файлове с файлов формат, който лесно да мигрира между различни типове операционни системи.

С оглед целта на научните изследвания в дисертационния труд и изброените по-горе критерии е избран двоичният вид на представяне на данни в обикновен низ от символи

поставен в обикновен текстов файл. Двоичният вид гарантира използването само на два вида символи – 1 и 0. Представянето в двоичен вид е заложено в настоящите компютърни системи. Конвертирането от и в двоичен вид на реалните данни – секретните ключове, е „позната“ и често използвана операция в една компютърна система. На практика генерирането на ключа може да се направи директно в двоичен вид. Двоичният вид е най-универсален вид на представяне на данни от страна на софтуера и хардуера.

Еднородна среда. Определение за еднородна среда: За постигане на целта на дисертацията и с оглед на избраните средства и методи като еднородна среда се определя среда, в която данните и околната им среда са представени с едни същи по вид, размер и брой символи с еднаква обща дължина.

Вида на символите според битовото представяне на данните е 1 и 0, а общата дължина със средата е подбрана да бъде 1024 символа. Всеки един символ е с размерност от 8 бита или 1 байт, а един текстови файл съхраняващ данните за даден ключ е 1 килобайт.

Средства и похвати за осъществяване на алгоритъма за поставяне на данни в еднородна среда. Алгоритъмът трябва да може да може да работи с гореописаните данни и средата, в която ще бъдат поставени. Освен това трябва да постига целта си – осигуряване на подходящ вид на данните за обучение и за работа на алгоритмите от машинно обучение. В тази връзка могат да се определят два основни етапа в работата му – преобразуването на данните в подходящ вид и поставянето им в еднородна среда. След осъществяването на тези два етапа е възможно и взаимодействието му с алгоритмите за машинно обучение посредством обработените вече данни.

Етап 1: Преобразуване в подходящ вид. Първата стъпка от работата на алгоритъма е преобразуването на данните в подходящ вид. Както беше пояснено по-рано, вида се подбира според целите и задачите, които трябва да се постигнат и решат от системата. За целите на изследванията в дисертацията са избрани данни от двоичен тип. Той гарантира използването само на два вида символи 1 и 0. Това от една страна определя лесното поставяне на данните в еднородна среда, а от друга благоприятства за използването на биполярна класификация на контролирано машинно обучение. Двата алгоритъма от машинно обучение (kNN, SVM), които се използват принадлежат към контролираното машинно обучение. Използването на биполярната класификация е предпоставка за постигането на по-добри крайни резултати.

Етап 2: Поставяне в еднородна среда. Средата, в която се поставят конвертираните данни е определяща за работата на алгоритъма. Тя трябва да съответства на данните и да осигурява достатъчен контраст, който е необходим и важен за работата на алгоритмите от машинното обучение.Съгласно приетата дефиниция за еднородна среда, това е среда, в която данните и околната им среда са представени с едни същи по вид, размер и брой символи с еднаква обща дължина. От изложеното до тук е видно, че „околна“ среда на данните са символите 1 и 0 или само единия от двата. От гледна точка на контраста използването на двата едновременно води до намаляване на контраста, а от там и до постигането на по-трудно различими данни. Изборът на един от двата варианта – 1 или 0 е равнозначен, а разликата между тях е в голямата си част субективна и в малка практична. Естествено 0 (нулата) се възприема, като начало, основа, както в електротехниката, а и в електрониката това е в буквален смисъл.

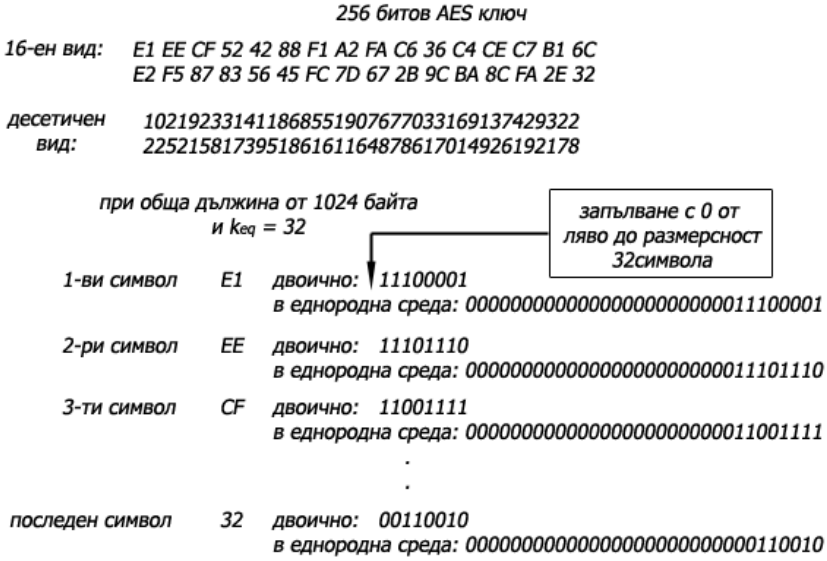
Друг фактор, с който трябва да се съобрази еднородната среда, е нейната дължина, която е важна заради зависимостта от нея на контраста на данните в тази среда. При малка

дължина съществува опасност от възникване на „препълване“, при което ключ не може да се постави в средата с цялата си дължина в двоично представяне. Възникването на препълване е неприемливо от гледна точка на решавания проблем. С оглед на избраните симетрични ключове и тяхната дължина, след предварителен анализ на различни варианти за дължина на средата изборът е за дължина от 1024 символа или 1024 байта, която е минималната необходима и достатъчна за постигане на целите. Дължината 1024 е число, което е симетрично на представените данни, а такава симетрия дава възможност и за бъдещо хардуерно реализиране на алгоритъма.

По този начин съгласно горните съображения в предложения алгоритъм всеки един символ на секретния ключ се представя в двоичен вид и се поставя в еднородна среда като се допълва с 0 от ляво до необходимата размерност. Допълването от ляво е възприето подобно на естествената посока на изписване на данните. Допълването е възможно да е също така от дясно или от двете страни. Конкретният избор за допълване само от ляво е въпрос на избор при интерпретирането на алгоритъма, тъй като няма конкретни съображения и контрастиращи разлики между трите варианта. Запълването с 0 се реализира посредством коефициент k_{eq} (55), чрез който се определя необходимото запълване до определената размерност. Във формулата L_e е приетата (определената) предварително дължина на еднородната среда, а S_{key} е размера на ключа в битове (bits), който се изравнява до байтове (bytes) чрез просто деление на 8.

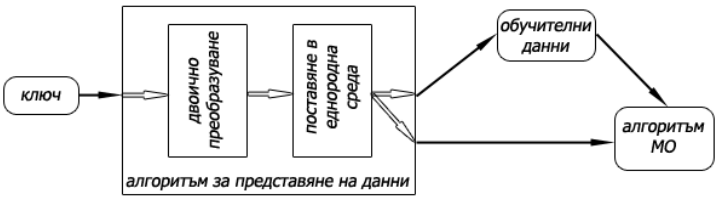
$$K_{eq} = \frac{L_e}{S_{key}/8} = 8\frac{L_e}{S_{key}} \tag{55}$$

Преобразуването и поставянето в еднородна среда на всеки символ е показано на Фиг.53. След преобразуването на всяко число в двоичен вид то се допълва с 0 от ляво до размерност 32. Като резултат в случая с примерния AES ключ се получават 32 части от по 32 елемента в обща размерност от 1024.



Фиг. 53 Допълване отляво до определена размерност

Взаимодействие на алгоритъма за поставяне на данните в еднородна среда с алгоритмите за машинно обучение. На Фиг.54 е представена обобщена схема на взаимодействие на алгоритъма с алгоритмите от машинно обучение.



Фиг. 54 Схема на взаимодействие

За определяне на сложността на предлагания алгоритъм посредством асимптотична нотация се използват функциите, които характеризират времето за изпълнение на алгоритмите. Определянето на сложността може да се приложи и за функции, които характеризират друг аспект на алгоритмите, например обем данни, които използват, или дори с функции, които не са директно свързани с алгоритмите [7]. Представянето на сложността на алгоритмите с Big O нотация се базира на математическа нотация, описваща ограничаващото поведение на функция, когато аргументът има тенденция към определена стойност или безкрайност. Според предложената теория [7] съпоставена към алгоритъма за поставяне в еднородна среда неговата сложност е $O(n^2)$. Ако се приеме, че за обхождането на данните е необходимо циклично действие n (най-трудоемкото) при преобразуването на данните в двоичен вид и за поставянето им в еднородната среда още едно циклично действие n, то математическата препратка е: $2n=O(n^2)$.

2.5 Изводи към втора глава

Основните изводи от изследванията, представени във втора глава на дисертационния труд мога да бъдат формулирани както следва:

1. Алгоритмите от машинно обучение kNN и SVM са подходящи за постигане на целта на дисертацията, защото са част от контролираното машинно обучение и са приложими за решаване на задачи за класификация.
2. Избраните алгоритми за симетрична криптография имат секретни ключове с различна дължина и характеристики, които отговарят на поставените изисквания за използване на машинно обучение за определяне на вида на ключа .
3. Избраният двоичен вид на представяне на данните, както за обучение, така и за работа на алгоритмите, дава възможност за използване на биполярна класификация, която съответства на възможностите за класификация с избраните алгоритми за машинно обучение. Освен това е лесна за интерпретация както на софтуерно, така и на хардуерно ниво.
4. Избраният тип файл за съхранение на тренировъчните данни е текстов файл, който възможно най-простият и лек начин както за съхранение на данните, така и за пренос.
5. Създаденият алгоритъма за поставянето на данни в еднородна среда е гъвкав и дава възможност за различни варианти на реализация относно запълването на околната среда и използването на различна дължина.
6. Оценката за сложността на предложения алгоритъм за представяне на данните позволява лесно реализиране в софтуерна среда и способства евентуална бъдеща хардуерна реализация.

Глава 3. Прилагане на алгоритми от машинно обучение и алгоритъм за поставяне в еднородна среда за разпознаване на секретни ключове.

Трета глава на дисертационния труд представя практично прилагане в програмен продукт (приложение) на избраните алгоритми за машинно обучение и предложеният алгоритъм за поставяне на данни в еднородна среда. Разделена е на три основни части. Първата определя изискванията към програмния продукт. Втората представя неговото проектиране, а третата програмната му реализация.

3.1 Определяне на изискванията към приложението за тестване на предложения подход и програмното осигуряване за неговата реализацията.

Изискванията към приложението са съобразени с всички фактори, които трябва да се вземат под внимание, като:

- Основната цел на дисертационния труд.
- Постигане на определените задачи.
- Провеждане на експерименти.

С оглед на тези фактори са формирани следните изисквания към приложението:

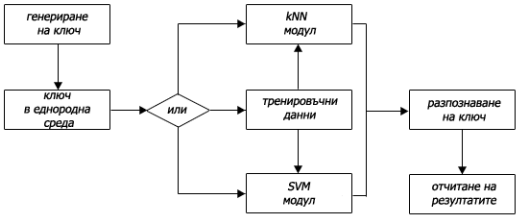
- Генериране на различни секретни ключове чрез определените за използване в дисертационния труд синхронни криптографски алгоритми;
- Имплементация на избраните алгоритми от машинно обучение;
- Подходяща обработка съгласно избрания вид на данните;
- Реализиране на еднородна среда и поставянето на данните в нея;
- Определяна на различен размер на еднородната среда;
- Създаване на тренировъчни данни според еднородната среда и вида на използвания криптографски алгоритъм, необходими за обучението на алгоритмите за машинно обучение;
- Разпознаване на вида на криптографския алгоритъм на конкретен секретен ключ с помощта на алгоритмите на машинно обучение.
- Възможност за различна интерпретация на алгоритмите;
- Възможност за провеждане на експериментални постановки и подходящо отчитане на резултатите съгласно определените критерии и параметри;
- Подходящ интерфейс и функционалност за използване в учебния процес при провеждане на учебни занятия по тематиката, както и за бъдещи научни изследвания.

3.2 Проектиране на приложението CryptoAndML.

Изграждането и разработката на приложението са съобразени с целта на дисертацията и формулираните задачи за изпълнение. В тази връзка са очертани няколко основни възли за реализация на необходимата функционалност на приложението: програмна реализация на генериране на секретен ключ за всеки един от криптиращите алгоритми; програмна реализация на обработка на данните в подходящия избран тип и поставянето им в еднородна среда; реализиране на модули за всеки един от двата алгоритъма на машинно обучение; модул за генериране и етикиране на тренировъчни данни, съобразени с размера на еднородната среда; модул за провеждане на единични и комплексни експерименти с отчитане на определените критерии и параметри; добре информиран и функционален

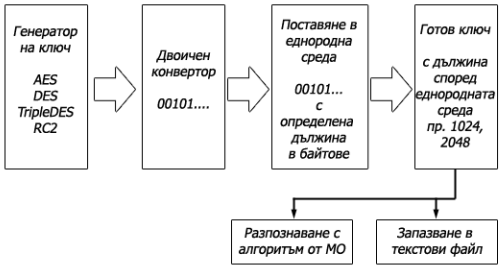
графичен потребителски интерфейс за да даде възможност за използването на продукта за учебни и научни цели.

На Фиг.55 е представена блокова схема на приложението. В алгоритъма в обобщен вид са показани основните модули, от които се състои приложението.



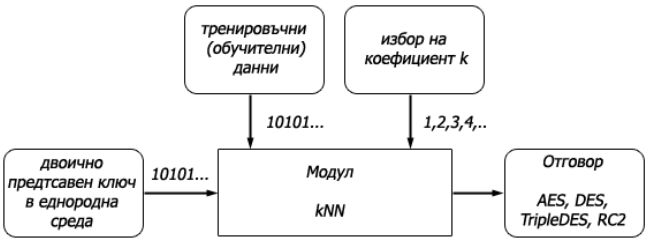
Фиг. 55 Схема на приложението „CryptoAndML“

Използването на софтуерното приложение започва с генериране на секретен ключ от даден криптографски алгоритъм, преобразуването му в двоичен вид и поставянето му в еднородна среда. За запазване и съхранение на тренировъчните данни в подходящ формат се следва алгоритмичната последователност, показана на Фиг.56.



Фиг. 56 Осигуряване на подходящ вид на данните

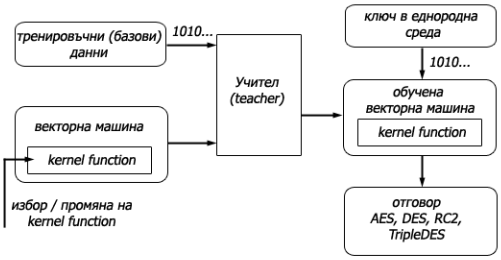
Два отделни модула реализират двата алгоритъма от машинно обучение – kNN и SVM. На Фиг.57 и Фиг.58 са показани проектираните схеми, съответно на алгоритмите kNN и SVM. За Алгоритъма kNN при използването му в режим на обучение като входни данни се подават набора тренировъчни данни и избран коефициент k, а в режим на обучение входът е двоично представения ключ в еднородна среда, а отговорът на класификатора е вида на подадения секретен ключ.



Фиг. 57 Схема на модул kNN

За Алгоритъма SVM при използването му в режим на обучение като входни данни се подават набора тренировъчни данни и избрана функция на ядрото, а в режим на обучение

входът е двоично представения ключ в еднородна среда, а отговорът на класификатора е вида на подадения секретен ключ.

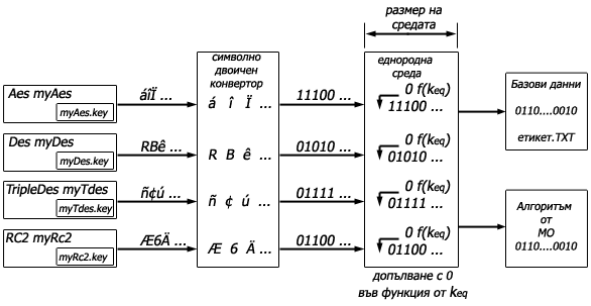


Фиг. 58 Схема на модул SVM

3.3 Програмна реализация на приложението.

Приложението е разработено на програмния език C# върху платформата .NET Framework 4.8, в средата за програмиране и дизайн на Visual Studio 2012 и 2019. Използвани са също и библиотеки от платформата Accord .NET Framework 3.8 за имплементиране и разработване на модулите за реализация на алгоритмите от машинно обучение. За реализацията са използвани всички възможности на обектно ориентираното и многонишково програмиране, тъй като алгоритмите за машинно обучение са изчислително сложни и изискват повече процесорно време [60], особено алгоритъма SVM.

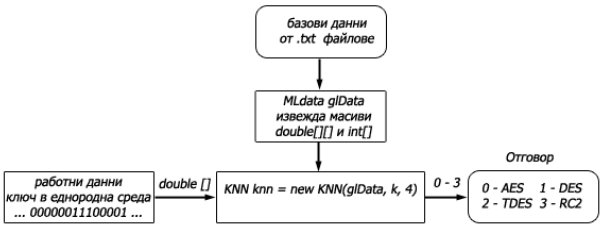
Четирите избрани синхронни криптографски алгоритми са вградени чрез платформата .NET Framework посредством наличните библиотеки и класове като се създава обект от съответния клас за имплементиране на дадения криптографски алгоритъм. В същото време се генерира и съответният криптографски ключ, който се разделя на отделни символи (размерността на един символ е 8 бита или 1 байт). Изчислява се коефициента необходим за поставяне на ключа в еднородна среда – k_{eq} , представен в израз (55). Този коефициент е в пропорционална зависимост от големината на еднородната среда и дължината на секретния ключ и съобразно с него всеки символ се допълва до определената дължина с 0 от ляво. Така представените данни на секретният ключ поставен в еднородна среда се подават на модулите на алгоритмите от машинно обучение или се подават към модула за създаване на базови (тренировъчни, обучителни) данни – Фиг.61.



Фиг. 61 Програмна реализация за осигуряване на подходящ вид на данните

Модула за имплементиране на алгоритъма от машинно обучение kNN е реализиран с помощта на библиотеките на платформата Accord.NET и специално създадени класове за целта. Тази реализация на алгоритъма се нуждае от определено представяне на входните (обучителни и работни) и изходните данни. Поради това са организирани допълнителни класове за трансформация и преразпределение на данните. На Фиг.62 е показана

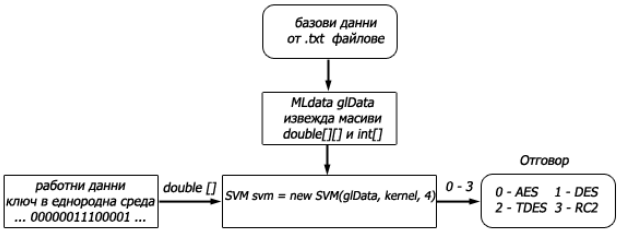
програмната реализация на модула. Конструктора за създаването на обекта kNN дава възможност за подаване на различен коефициент k и за определяне на броя класове на подаваните данни. Освен това този обект се създава и работи в условията на друга нова нишка, различна от основната на приложението. Това се налага поради изчислителната сложност и нужното процесорно време на алгоритмите от машинно обучение и за използване на многонишковата архитектура на съвременните компютърни системи.



Фиг. 62 Реализация на модул kNN

Реализацията на модула за алгоритъма SVM е с помощта на библиотеките на платформата Accord.NET и специално създадени за целта класове. Тук също се изисква определено представяне на входните (обучителни и работни) и изходните данни. Допълнителното е, че при този алгоритъм се създава векторна машина, която се обучава от „учител“ и след това се работи с нея.

Докато при алгоритъма kNN се изисква един цикъл на обръщения към процесора, то при алгоритъма SVM са необходими два. Освен това този алгоритъм използва по-сложни математически функции за „функциите на ядрото“. Поради тези причини освен отделянето на работата на алгоритъма в отделна нишка е направена и организация за следене на работата му: следи се обучението на векторната машина. След като се обучи машината тя се използва до края на изпълнението на цялото приложение, с което се спестява по един от циклите на обръщения към процесора и се намалява времето за изпълнение средно с около 15% (в зависимост от хардуера на системата, върху която се изпълнява). На Фиг.63 е показана реализацията на алгоритъма SVM. Обекта svm от класа SVM изпълнява главната роля. Реализиран е така, че да има възможност за определяне на различна функция на ядрото на векторната машина. Освен това поддържа и определяне на броя класове, с които се работи.



Фиг. 63 Реализация на модул SVM

3.4 Изводи към трета глава.

Основните изводи от изследванията, представени в трета глава на дисертационния труд мога да бъдат формулирани както следва:

1. Представените програмни инструменти за проектиране и имплементация на представените алгоритми спестяват време за написване на повтарящ се код и позволяват да се фокусира реализацията върху самия модел, а не върху обкръжаващата го програмна среда и нейните компоненти;
2. Избраните програмни инструменти позволяват използване на моделите върху голям набор от хардуерни устройства, в това число компютърни системи с един или няколко графични процесора, високо-производителни компютърни системи, изградени от голям брой специализирани машини и графични процесори;
3. При изпълнението си двата алгоритъма от машинно обучение изискват повече ресурс, поради което е предвидена и имплементирана многонишкова реализация;
4. По възможност е препоръчително запазването на състоянието на вече обучената векторна машина;
5. Алгоритъма за предварителна обработка на данните и поставянето им в еднородна среда е лесен за реализация. Имплементацията му е възможна в почти всеки от най-широко използваните езици от високо ниво;
6. Изборът на подходящ размер на еднородната среда е важен и зависим от входните данни, с които ще работи.

Основните резултати от изследванията, представени в трета глава на дисертационния труд мога да бъдат обобщени както следва:

1. Избрани са програмни инструменти, които предоставят необходимите средства за проектиране, имплементиране, настройване, тестване и използване на представените алгоритми и модели;
2. Представен е подход за подготовка, създаване, обработка и анализ на криптографски данни за работа в средата на алгоритми за машинно обучение;
3. Описани са спецификите при проектирането, конфигурирането и създаването на необходимите слоеве и класове на представения модел на съвместна работа на разглежданите алгоритми за класификация на криптографски данни;
4. Представеният подход позволява проектиране, конфигуриране и имплементиране на машинно обучение в криптографски системи, чрез подходяща предварителна обработка на данните, както за експериментални изследвания, така и разработване на продуктови програмни системи.

Глава 4. Планиране на опитната постановка. Провеждане и анализ на експериментални резултати.

Четвърта глава на дисертационния труд представя планирането и провеждането на експерименти за оценка на приложението на предложения алгоритъм за поставяне на данни в еднородна среда и оценка на взаимодействието му с алгоритмите за машинно обучение.

4.1 Планиране на експериментална постановка.

Експерименталната постановка е планирана да провери и докаже възможностите за разпознаване на вида на алгоритъма на секретния ключ чрез машинно обучение. Чрез разпознаване може да се увеличи устойчивостта на симетричните криптографски алгоритми. Осигуряването на коректно разпознаване на секретния ключ може да осигури процес на криптиране – декриптиране с повече от един синхронен криптографски алгоритъм, което води до по-голяма устойчивост на атаки по метода на грубата сила.

Провеждането на различните експерименти се извършва със създаденото за целта приложение за генериране на случайни ключове, поставянето им в еднородна среда и тяхното разпознаване с алгоритми за машинно обучение. С негова помощ се изследва работата на алгоритъма за поставяне на данни в еднородна среда и взаимодействието му с алгоритмите на машинно обучение. Целта е да се оцени влиянието на основните параметри за работата на една такава система – размера на еднородната среда, дължината на секретния ключ за разпознаване, размера (броя) на базовите данни, различните варианти на използваните алгоритми за машинно обучение и хардуера върху, който се използват. Задължително е отчитането и на времето за изпълнение на задачите, за да се даде отговор на възможността за практическа реализация в общодостъпни компютърни системи.

4.2 Планиране на експериментална оценка.

Планирането на експериментите за оценка на приложимостта на предложения алгоритъм за поставяне на данни в еднородна среда и взаимодействието му с алгоритмите за машинно обучение включва използването на различни конфигурации за следните параметри:

Входни данни. Използват се четири алгоритъма за симетрично криптиране – AES, DES, TripleDES и RC2, чрез които се генерират четири различни по вид секретни ключове с различна дължина 64, 128 и 256 бита. Това са входните данни за системата.

Еднородната среда. Големината (размера) на еднородната среда е ключова за работата на алгоритъма за поставяне на данни в такава среда. Тя е в зависимост от размера на входните данни. За експериментална оценка на влиянието на дължината на еднородната среда се използват три различни размера – 1024, 1536 и 2048 символа.

Базови данни. Важна част за работата на всеки един алгоритъм от контролирано машинно обучение са базовите (етикираните, обучаващи) данни. За експериментална оценка на влиянието на обема на базовите данни се използват опции с 500, 1000 и 1500 примера (базови данни) за всеки един вариант от входните данни.

Алгоритми за машинно обучение. За експериментална оценка на влиянието на параметрите на алгоритъма kNN се използват няколко стойности за коефициента k – 3, 29, 43, 53 и 63, а за алгоритъма SVM – варианти с линейна и полиномиална функция на ядрото.

Хардуер. Използваните алгоритми са изчислително сложни и изискват значително процесорно време, като това особено важи за алгоритмите за машинно обучение. За експериментална оценка на влиянието на хардуерната система са използвани няколко хардуерни системи от различно поколение.

Изброените опции позволяват провеждането на експериментална оценка с 480 различни комбинации за конфигурациите на различните параметри. От всички възможни комбинации са избрани 4 експеримента с конкретни конфигурации, за да се проверят и докажат възможностите за разпознаване на вида на алгоритъма на секретния ключ чрез машинно обучение, което да даде възможност за увеличаване на устойчивостта на синхронните криптографски алгоритми. Общия брой на проведените опити с различни конфигурации е 26.

Експеримент 1. Целта на този експеримент е да се определи подходящия размер на еднородната среда и да се докаже зависимостта му от размера на входните данни.

Експеримент 2. Целта на този експеримент е да се изследва зависимостта на системата от алгоритъма за поставяне на данни в еднородна среда и използваните алгоритмите за машинно обучение от броя на известните примери (базови данни).

Експеримент 3. Целта на този експеримент е да се оцени влиянието на различните варианти на използваните алгоритми за машинно обучение върху работата на системата.

Експеримент 4. Целта на този експеримент е да се оценят резултатите на системата от алгоритъма за поставяне на данни в еднородна среда и използваните алгоритмите за машинно обучение при изпълнение върху различни хардуерни конфигурации.

За оценка на получените експериментални резултати се оценява поведението на обучения модел по отношение на точността и грешката по време на изпълнение. Точността определя колко често моделът е бил коректен при класификацията.

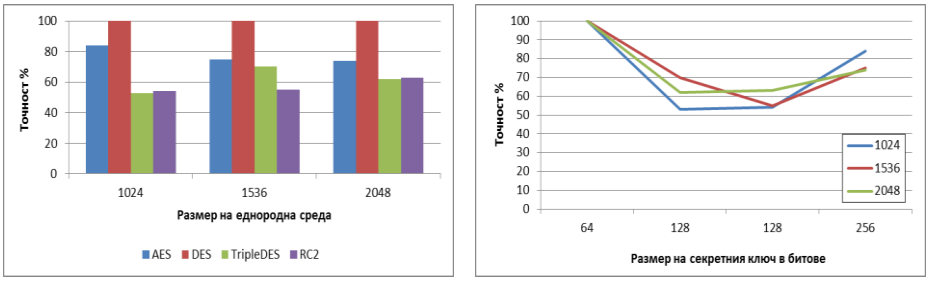
Оценка на комплексната работа и реалната „мощност“ на обучения модел се реализира чрез отношение на точността спрямо „изразходваното“ време. Мощността определя силата на модела и се базира на коректно свършена работа за време в минути.

Така получените обобщени резултати от експериментите се съпоставят с описаните реализации в раздел 1.4.1 и 1.4.2 за разпознаване на секретен ключ базирано на електромагнитен анализ на хардуерна реализация на криптографски алгоритми и извлечени от IPT (Intel Processor Trace) следи. Целта на това сравнение е да се определят прилики и разлики в двата подхода на разпознаване поради факта, че се използват едни и същи криптографски алгоритми като обект и еднотипни алгоритми от машинното обучение като средство.

4.3 Експериментални резултати.

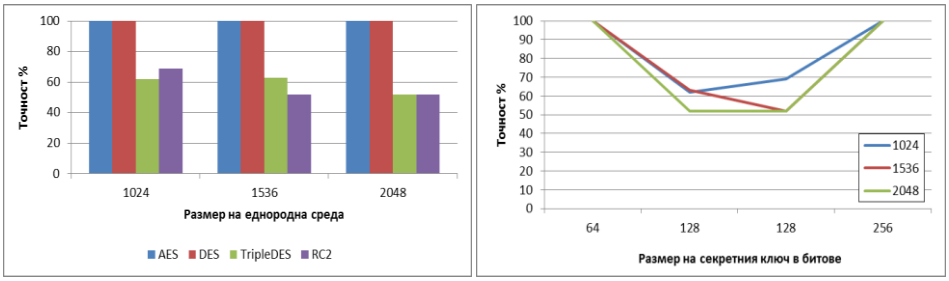
Експеримент 1. Първият експеримент има за цел да оцени влиянието на размера на еднородната среда върху работата на системата и в частност резултатите от класификацията. Базовите данни са с по 1000 примера за всеки от четирите алгоритъма, използвани за генериране на секретни ключове. Входните данни са по 100 генерирани секретни ключа за всеки от криптиращите алгоритми.

На Графика 1 са показани детайлни резултати съпоставени едни към други за различен размер на еднородната среда при работа с алгоритъма kNN.



Графика 1: Точност при модел с kNN и различна дължина на еднородната среда

На Графика 2 са показани детайлни резултати съпоставени едни към други за различен размер на еднородната среда при работа с алгоритъма SVM.



Графика 2: Точност при модел с SVM и различна дължина на еднородната среда

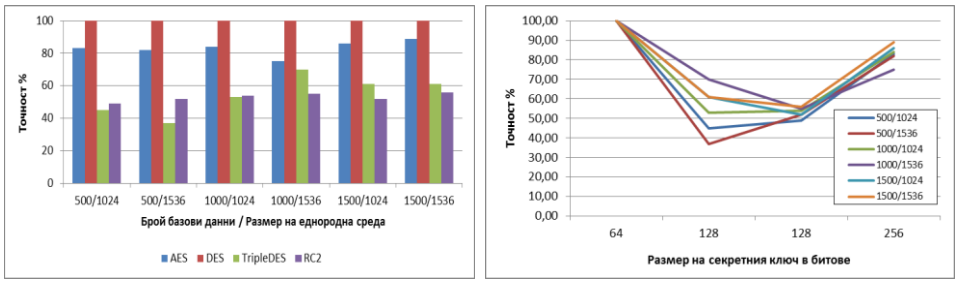
На база на получените експериментални резултати от експеримента могат да бъдат направени следните изводи:

• Резултатите са с висок процент на коректно разпознаване в двете крайности на размера на разпознавания секретен ключ, което е очакван резултат с оглед на по-добрия контраст на данните поставени в еднородна среда. Размера на еднородната среда при тях почти няма никакво значение.

• При разпознаването на еднакви по дължина ключове от 128 бита резултатите за коректно разпознати ключове са в рамките на 55-75%, като влиянието на размера на еднородната среда е ясно изразено. Резултатите показват, че при голям размер на средата се стига до размиване на данните поставени в нея, а при малък размер се стига до препълване. Освен това по-големият размер води до повече необходимо време за изпълнение.

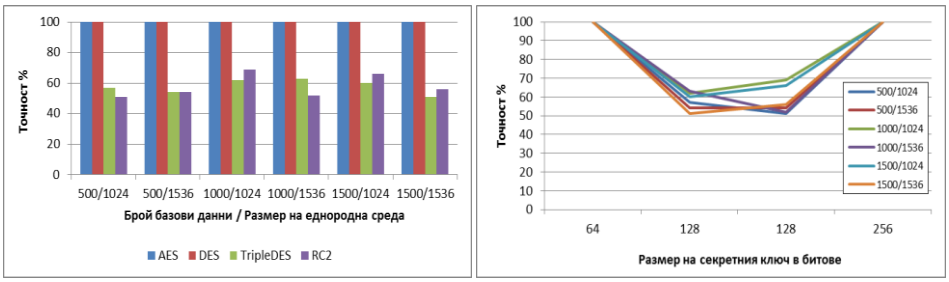
Експеримент 2. Вторият експеримент има за цел да се изследва зависимостта на системата от алгоритъма за поставяне на данни в еднородна среда и използваните алгоритмите за машинно обучение от броя на известните примери (базови данни). Входните данни са по 100 генерирани секретни ключа за всеки от криптиращите алгоритми. Направени са опити с 500 и 1500 базови данни. Използват се и резултатите от опитите с 1000 базови данни от експеримент 1.

На Графика 3 са показани детайлни резултати съпоставени едни към други за различен размер на еднородната среда и брой базови данни при работа с алгоритъма kNN.



Графика 3: Точност при модел с kNN и различен брой базови данни

На Графика 4 са показани детайлни резултати съпоставени едни към други за различен размер на еднородната среда и брой базови данни при работа с алгоритъма SVM.



Графика 4**:** Точност при модел с SVM и различен брой базови данни

На база на получените експериментални резултати от експеримента могат да бъдат направени следните изводи:

• Резултатите са с висок процент на коректно разпознаване в двете крайности на размера на разпознавания секретен ключ, дължащо се на по-добрия контраст на данните поставени в еднородна среда.

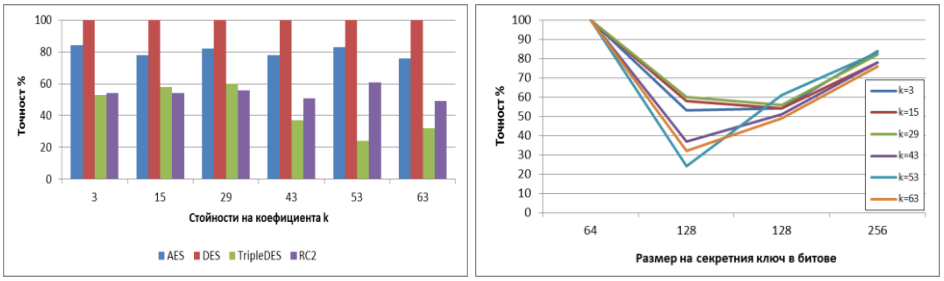
• При разпознаването на еднакви по дължина ключове от 128 бита, резултатите са по-нисък процент на разпознаване в сравнение с експеримент 1. Точността при 500 броя на базовите данни намалява, което показва, че този обем на базовите данни не е достатъчен за минимално базово осигуряване с тренировъчни данни на алгоритмите за машинно обучение.

• Времето за изпълнение значително нараства при 1500 броя базови данни и размер на еднородната среда от 1536 байта (символа). Това показва, че големия брой базови данни не води до задължително подобрение на резултатите, а увеличаването на размера на еднородната среда без причина е неоснователно. Всичко това съпоставимо с времето за изпълнение е много ясно изразено.

• Получените резултати са най-добри при размер на еднородната среда от 1024 байта и при 1000 и 1500 броя на базовите данни, но при 1000 на брои базови данни времето за изпълнение е по-малко, което налага извод за оптимален вариант при размер на еднородната среда от 1024 байта и 1000 броя базови данни.

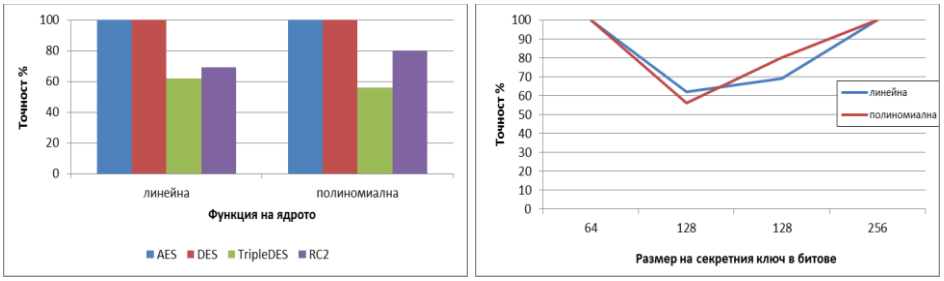
Експеримент 3. Третият експеримент има за цел да провери влиянието на различните варианти на използваните алгоритми за машинно обучение върху работата на системата. Входните данни са по 100 генерирани секретни ключа за всеки от криптиращите алгоритми. Направени са опити с по 1000 броя базови данни. Вариантите на алгоритъма kNN са при различни стойности на k – 3, 15, 29, 43, 53 и 63, а за SVM вариантите са при линейна и полиномиална функция на ядрото. Използват се и резултатите от опитите с 1000 базови данни от експеримент 1.

На Графика 5 са показани детайлни резултати съпоставени едни към други за различен коефициент k на алгоритъма kNN.



Графика 5: Точност при модел с kNN с различен коефициент k

На Графика 6 са показани детайлни резултати съпоставени едни към други за различен размер на еднородната среда и брой базови данни при работа с алгоритъма SVM.



Графика 6: Точност при модел с SVM с различна функция на ядрото

На база на получените експериментални резултати от експеримента могат да бъдат направени следните изводи:

• Резултатите са с висок процент на коректно разпознаване при големия контраст на данните – за 64 и 256 битов ключ, като при SVM точността на разпознаване е 100% ;

• Резултатите показват по-висок процент на точността на разпознаване при коефициент k със стойност 3. Високи стойности за точността на разпознаване се постигат и при стойности за k 15 и 29, което подкрепя тезата, че стойноста на коефициента k трябва да е по-малка от корен квадратен от броя на примерите;

• При двата варианта на функция на ядрото при SVM точността на разпознаване е 100% за 64 и 256 битов ключ. При двата 128 битови ключа „полиномиалната“ функция показва по-добри резултати, но в тези случаи се увеличава необходимото време за изпълнение, което е значително по-голямо при разпознаване с алгоритъм SVM в сравнение с алгоритъм kNN;

Експеримент 4. Четвъртият експеримент има за цел да се оцени резултатите на системата при работа върху различни хардуерни конфигурации. Входните данни са по 100 генерирани секретни ключа за всеки от криптиращите алгоритми. Направени са опити с по 1000 броя базови данни и размер на еднородната среда 1024 байта. Използвани са коефициент k със стойност 3 за алгоритъм kNN и линейна функция на ядрото за алгоритъм SVM. Използвани са следните четири хардуерни конфигурации:

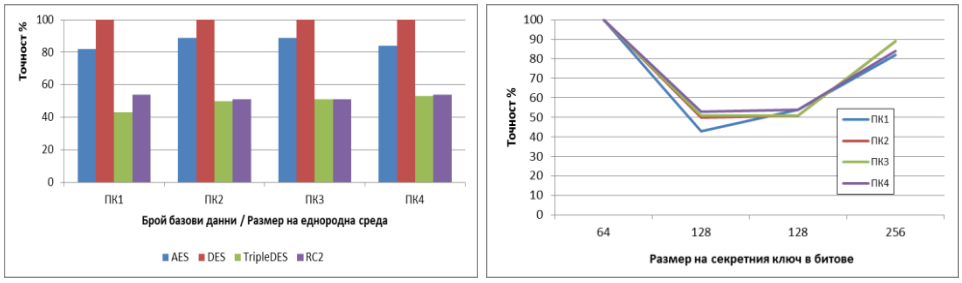
• ПК1 - Intel Xeon E5450/4 cores/3.00GHz, 8GB/DDR2-400Mhz, HDD, Windows 7 Ultimate/x64/SP1;

• ПК2 - Intel Xeon E5450/4 cores/3.00GHz, 8GB/DDR2-400Mhz, SSD, Windows 7 Ultimate/x64/SP1;

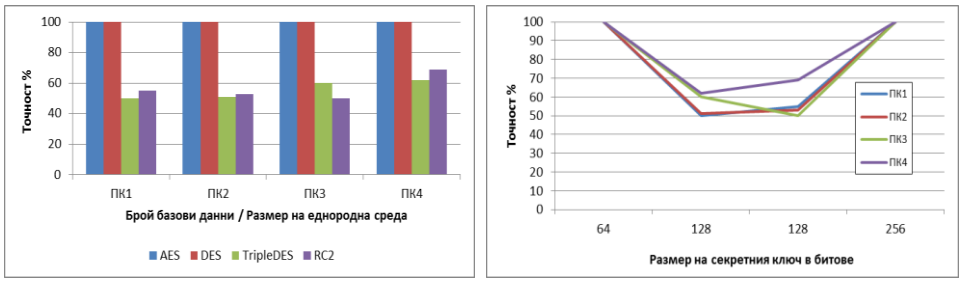
• ПК3 - Intel Core i5-4460/4 cores/4 threads/3.2Ghz/MaxTF 3.4GHz, 8GB/DDR3-1600MHz, SSD, Windows 10 Pro/x64/SP1;

• ПК4 - Intel Core i5-11400H/6 cores/12 threads/2.7Ghz/MaxTF 4.5GHz, 16GB/DDR4-3200MHz, NVMe SSD, Windows 10 Pro/x64/SP1;

На Графика 7 и 8 са показани детайлни резултати съпоставени едни към други за различни хардуерни конфигурации при използване на алгоритъма kNN и SVM.



Графика 7: Точност при модел с kNN при различни хардуерни конфигурации



Графика 8: Точност при модел с SVM при различни хардуерни конфигурации

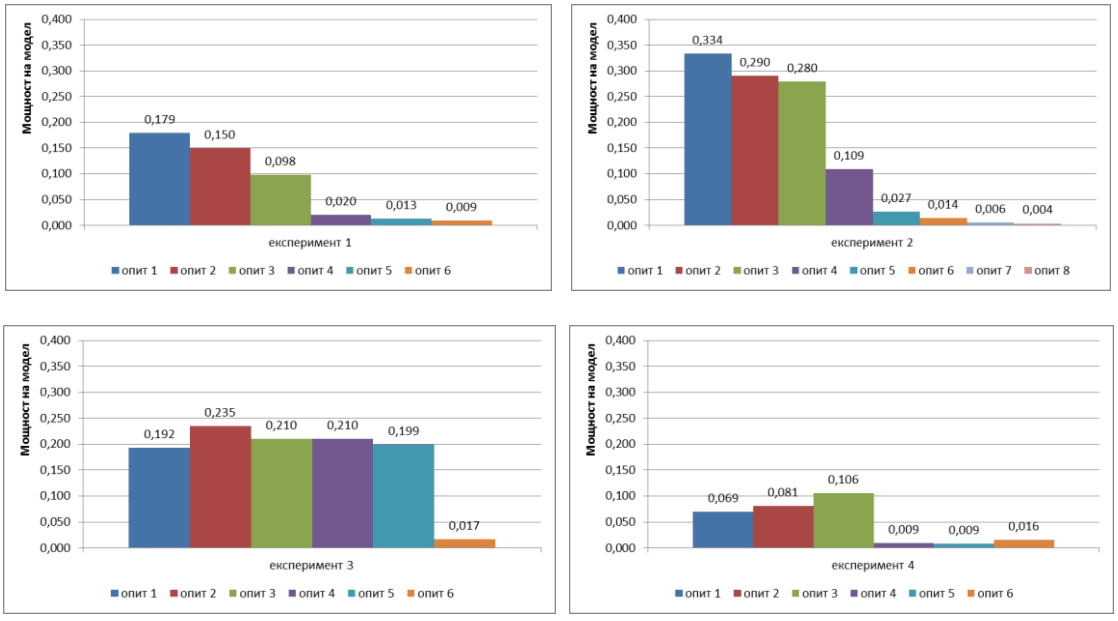
На база на получените експериментални резултати от експеримента могат да бъдат направени следните изводи:

• Резултатите за точността за коректно разпознаване са с около 14% по-добри при най-новата компютърна система с процесор от последно поколение (ПК4) в сравнение с по-старите конфигурации. Основна причина, за което се явява Intel® Deep Learning Boost (Intel® DL Boost) - вградена процесорна технология, предназначена да ускорява случаите на използване на задълбочено обучение. Тя разширява Intel AVX-512 с нова векторна инструкция за невронна мрежа (VNNI), която значително увеличава производителността на изводите за дълбоко обучение спрямо предишните поколения;

• Разликите при точността за коректното разпознаване между различните компютърни конфигурации са малки, но за сметка на двойно (при kNN) и тройно (при алгоритъма SVM) увеличение на времето, необходимо за изпълнение при по-старите системи;

• Точността при използване на алгоритъма SVM е голяма, но поради тежестта си коефициента му на „мощност“ е по-малък. Видно е, че с увеличаване на производителността на компютърните технологии този недостатък може бъде преодолян (Графика 9);

На Графика 9 са показани коефициентите на мощност на всички опити проведени в рамките на четирите експеримента.



Графика 9: Коефициентите на мощност на всички опити

Резултатите показват, че в 5 от 7-те най-добри резултати, размерът на еднородната среда е 1024 байта (символа). Като се вземат предвид използваните размери на ключове може да се определи, че размера на еднородната среда е в следната зависимост от големината на най-големия ключ:

$$L_{he} = l_{max}^2 \tag{59}$$

Тук L_{he} е размера на еднородната среда в байтове, а l_{max} е размера на най-големия секретен ключ в байтове, който се използва.

4.4 Сравнителен анализ на получените резултатите

В раздел 1.4.1 е разгледано разпознаване и възстановяване на секретен ключ, реализирано посредством електромагнитен анализ на хардуерна реализация на криптографски алгоритми AES, DES и RSA и извличане на характеристики с помощта на подход с машинно обучение, което ще наричаме подход-1. Спецификата на подхода е, че се използват седем бита от секретният ключ за разпознаване – 2, 3, и 4 бит, първи се игнорира, а 5, 6, 7 и 8 им се задава стойност „0“. Проведени са различни по фактори експерименти и са декларирани обобщени резултати, представени в Таблица 14[85].

Таблица 1: Експериментални резултати за подход-1

	2 бит		3 бит		4 бит	
	коректно разпознати, %	време, сек.	коректно разпознати, %	време, сек.	коректно разпознати, %	време, сек.
Support Vector Machine (SVM)	45,70	0,35	49,30	0,34	55,40	0,30
Random Forest (RF)	58,00	0,63	56,90	0,64	79,20	0,74
Naive Bayes (NB)	55,70	0,06	57,00	0,07	52,50	0,08

В раздел 1.4.2 е представен метод за идентифициране на симетрични криптографски алгоритми – AES, BF, CAST, DES, DES3, IDEA, RC2, RC4, SEED, чрез прилагане на конволюционна невронна мрежа към следите, извлечени от IPT (Intel Processor Trace), който ще наричаме подход-3. Резултатите от обучението са класифицирани в девет класа със 100% точност. Друг набор от данни съдържа следи, извлечени от различните симетрични алгоритми с различна дължина на секретния ключ. Резултатите от обучението са класифицирани в 36 класа с точност 70.55%.

Съпоставянето на резултатите се базира на два параметъра – точност на разпознаване и общо време за изпълнение, които са основни за оценяване на ефективността на всеки от следните сравнявани подходи:

– **Подход-1**: разпознаване реализирано посредством електромагнитен анализ на хардуерна реализация на криптографски алгоритми и извличане на характеристики с помощта на подход с машинно обучение;

– **Подход-2**: е предложения в дисертацията подход за разпознаване на симетричен секретен ключ чрез поставяне на данните в еднородна среда и класифициране с помощта на алгоритми от машинно обучение;

– **Подход-3**: с метод за идентифициране на симетрични криптографски чрез прилагане на конволюционна невронна мрежа към следите, извлечени от IPT.

На базата на получените в дисертационния труд експериментални резултати и публикуваните резултати за подход-1 и подход-3 може да бъде направен следния анализ на резултатите:

- ***Точност на разпознаване:***
  - Точността на правилно разпознатите секретни ключове посредством подхода предложен в дисертацията (**подход-2**) е по-висока в сравнение с подход-1 и подход-3, като в почти всички проведени експерименти точността с **подход-2** е **100%** за секретни ключове на AES и DES, като в някои експерименти с

вариране на различни конфигурационни параметри намалява до **88-90%** (Графика 10);

- При подход-1 точността на разпознаване е в границите от 55% до 63%, с едно пиково постижение от 79.2% (Таблица 15);
- И двата подхода подход-1 и **подход-2** използват алгоритъма SVM, като сравнението на постигнатата мощност от SVM показва **1.7 пъти по-добри резултати за подход-2** спрямо подход-1;
- Авторите на подход-1 докладват 87% коректно разпознаване за SVM при 256 битов ключ на AES, докато **при използване на подход-2 точността се увеличава с около 10-15%**;
- При подход-3 са декларирани 100% точност при първия набор алгоритми, която е еквивалентна на резултатите при **подход-2** с някои изключения при различни конфигурации на използваните параметри, за които точността е **88-90%**;
- При втория набор от алгоритми при подход-3, които са с различна дължина на секретния ключ, е декларирана точност 70.55% , докато при **подход-2** математически осреднения процент точност е **82.77%**.

- ***Време за изпълнение:***
  - Времето за изпълнение при подход-1 е в диапазона от 0.3 до 0.8 sec за SVM и RF, а при NB – средно около 0.07 sec за разпознаването на един бит;
  - При **подход-2** времето, необходимо за разпознаване на секретен ключ е в диапазона от **6.8** до **7.4 sec** за най-добрите постигнати експериментални резултати, получени при еднородна среда с размер от 1024 байта;
  - При ключ с дължина от 256 при **подход-2** средното време за изпълнение е **0.026 sec**, което означава, че и по този параметър на сравнение **подход-2** е с **по-добри показатели** от подход-1;
  - При подход-3 не са публикувани данни за време за изпълнение, но самата схема на изпълнение, която е свързана с извличане на данни от една компютърна система и предаването им на друга, която в последствие ги конвертира в изображения преди да ги анализира и класифицира, обуславя увеличение на времето за изпълнение в пъти спрямо **подход-2**.

На Таблица 15 са обобщени резултатите по критерий – точност на разпознаване.

Таблица 15: Сравнителна таблица на точността на сравняваните подходи

	Подход 1	**Подход 2**	Подход 3
максимална точност в %	79,2	**100**	100
минимална точност в %	55	**88**	70,55

Всеки от сравнените подходи има своите предимства и недостатъци. В Таблица 16 са показани някой от най-съществените прилики и разлики между сравняваните подходи. Като обща черта може да се определи предварителната обработка на входните данни. Съществената разлика се определя от допълнителното филтриране на данните, необходимостта от допълнително хардуерно осигуряване и вида на взаимодействие с входните данни при подход-1 и подход-3. **Подход-2** е в пряко взаимодействие с входните данни.

Таблица 16: Основни прилики и разлики между сравняваните подходи

	Подход 1	**Подход 2**	Подход 3
Пряко взаимодействие с входните данни	Не	**Да**	Не
Косвено взаимодействие с входните данни	Да	**Не**	Да
Предварителна обработка на входните данни	Да	**Да**	Да
Филтриране на данните	Да	**Не**	Да
Допълнително хардуерно оборудване	Да	**Не**	Да
Допълнителен високо производителен графичен процесор	Не	**Не**	Да
Лесен за разбиране и внедряване	Не	**Да**	Не

Проведените експериментални изследвания и получените от тях резултати, както и направеният сравнителен анализ на трите подхода подкрепят изследователската хипотеза на научните изследвания в дисертационния труд и потвърждават, че **подход-2**, предложен в дисертацията, дава възможност за решаване на проблеми в синхронната криптография с използване на алгоритъма за поставяне на данни в еднородна среда и класификация с машинно обучение.

4.5 Изводи към четвърта глава

Основните изводи от изследванията, представени в четвърта глава на дисертационния труд могат да бъдат формулирани както следва:

1. Алгоритмите за контролирано машинно обучение могат да се използват за решаване на проблеми с устойчивостта на симетричните криптографски алгоритми.
2. Предложеният алгоритъм за поставяне на данни в еднородна среда осигурява подходящ вид на обучаващите данни за алгоритмите от машинното обучение.
3. Големината на еднородната среда оказва влияние на контраста на двоичния отпечатък на секретния ключ, който се поставя в нея.
4. Точността на разпознаване на секретния ключ от алгоритмите на машинното обучение са в пряка зависимост от правилния подбор на големината на еднородната среда спрямо битовата дължина на секретните ключове на криптографските алгоритми, които се разпознават. Ако големината е по-малка от големината на използваните секретни ключове се стига до загуба на данни, а при прекалено голям размер на еднородната среда спрямо размера на секретните ключове е възможно прекалено голямо разреждане на двоичния отпечатък на ключа. И в двата случая се влошава работата на алгоритмите на машинното обучение.
5. Получените резултати от подхода за разпознаване на криптографските алгоритми от алгоритмите на машинното обучение с помощта на алгоритъма за поставяне на данни в еднородна среда, представен в дисертацията, са по-добри спрямо резултатите от подходите на подобните решения за разпознаване на секретен ключ, с които са сравнени. Подходът представен в дисертацията е лесен за реализиране и последващо използване.

Заключение

С развитието на компютърните технологии много от криптографските алгоритми губят част от своята сигурност. Скоростта на симетричните криптографски алгоритми е безспорно най-доброто им качество. Бързият възход на технологиите обаче дава тласък и отваря нови възможности за други научни направления. С оглед на това машинното обучение получава много нови възможности. В резултат на проведените научни изследвания се потвърждава работната хипотеза на дисертационния труд за приложимостта на методи и алгоритми за машинно обучение за създаване и обучение на модели за класификация и разпознаване на секретни ключове от симетрични криптографски алгоритми. Предложеният алгоритъм за поставяне на данни в еднородна среда е от важно значение за създаване на модели, които да подпомогнат обучението и подготовката на входните данни на алгоритмите за машинно обучение. С използване на предложените методи и алгоритми за отделните етапи за подготовка и обработване на входните данни (секретни ключове) се дават нови възможности на симетричните криптографски алгоритми и на сигурността на информацията, като цяло. Проведените експериментални изследвания и получените резултати за изследваните модели показват, че системни архитектури с използване на машинно обучение са подходящи за анализ и класификация на криптографски данни. С използване на представеният подход за проектиране, конфигуриране и имплементиране на машинно обучение с използване на алгоритъма за поставяне на данни в еднородна среда може да се осигури анализ и класификация на криптографски данни и в частност секретни криптографски ключове за постигане на коректното им разпознаване. Това веднага определя възможността за криптиране-декриптиране с повече от един вид симетричен алгоритъм в една комуникационна сесия. С оглед на това се увеличава устойчивостта на външни атаки на такива криптографски системи. Теоретично то увеличение на устойчивостта при атака от вида на „грубата сила“ е 4 пъти, ако се използват 4 различни криптографски алгоритъма.

Основните насоки за бъдещи научни изследвания в областта на анализ и класификация на криптографски данни чрез машинно обучение с използване на алгоритъма за поставяне на данни в еднородна среда са следните:

- Анализ и класификация на шифротекстове с цел разпознаване на криптиращия алгоритъм;
- Предложеният подход за имплементиране на машинно обучение в криптографски системи с алгоритъм за поставяне на данни в еднородна среда дава висока точност при класификацията и може да се използва за изграждане на реален продукт, достъпен за използване, като класификатор в модел на многопрофилно криптиране или криптиране с различни алгоритми;
- Дава се средство за постигане на симетрична криптографска комуникация без „уговорка“ между подател и получател за вида на ключа, който да се използва за криптиране и декриптиране, както и възможност да отпадане необходимостта от съхранение на секретен ключ за дадена комуникация;

НАУЧНИ, НАУЧНО-ПРИЛОЖНИ И ПРИЛОЖНИ ПРИНОСИ

Научни приноси:

- Предложен е алгоритъм за поставяне на данни в еднородна среда, осигуряващ подходящ вид на входните данни за алгоритмите на машинно обучение;
- Предложен е подход за формиране на обучаващо множество данни за класифициране на криптографски данни с помощта на предложения алгоритъм за поставяне на данни в еднородна среда;

Научно-приложни приноси:

- Предложен е подход за проектиране, конфигуриране и имплементиране на предложения алгоритъм за поставяне на данни в еднородна среда и алгоритми от машинно обучение за разпознаване на криптографски данни (Подход 2);

Приложни приноси:

- Предложеният алгоритъм за поставяне на данни в еднородна среда е реализиран заедно с алгоритми за машинно обучение kNN и SVM в система за разпознаване на криптографски данни;
- На базата на експериментални изследвания са определени подходяща конфигурация и параметри на предложения подход за разпознаване на криптографски данни с цел увеличаване на устойчивостта на симетричните криптографски алгоритми. Определен е размера на еднородната среда в зависимост от дължината на симетричния ключ (59);
- Реализираната система за разпознаване на криптографски данни и извършените експериментални изследвания позволяват реализирането на модел на многопрофилно криптиране или криптиране с различни алгоритми в единна среда.

СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ДИСЕРТАЦИОННИЯ ТРУД

1. D. Todorov, I. Penev, Hand-writen Digit Recognition by Support Vector Mashines, International scientific-practical conference of young scientists, graduate students and students, Kharkiv, Ukraine, 09-10 July 2017, UDC 004.55, pp. 122, УДК 681.518.54;
2. I. Penev, M. Karova, D. Todorov, Handwritten Digit Recognition Using Blob Detection and Machine Learning, Advances in Neural Networks and Applications'2018, St. Konstantin and Elena Resort, Bulgaria, 15-17 Sept. 2018, ISBN: 978-3-8007-4756-6 (**индексирана в Scopus**).
3. D. Todorov, Steganographic Embedding of Information in an Image Using a Template Matrix Depending on The Length of The Message, Computer Science and Technologies Journal, vol. 1, pp. 23, 2020, ISSN 1312-3335.
4. D.Todorov, M. Karova, Machine Secret Key Recognition in a Homogeneous Environment, International Conference Automatics and Informatics'2021, Varna, Bulgaria, 30 Sept.-2 Oct. 2021, ISBN: 978-1-6654-2661-9 (**индексирана в Scopus**).
5. D. Todorov, Zh. Zheynov, H. Valchanov, M. Karova, V. Aleksieva, I. Penev, A. Hakka, G. Spasova, I. Boychev, P. Genchev, G. Marinova, D. Dinev, P. Edreva, Investigation of the influence of a template matrix on the embedding of information in an image, Annual Journal of Technical University of Varna, ISSN 2603-316X (под печат).
6. D.Todorov, M. Karova, Appropriate Conversion of Machine Learning Data, Annual Journal of Technical University of Varna, ISSN 2603-316X (под печат).

Благодарности на:

Изказвам благодарности на доц. д-р инж. Милена Карова, доц. д-р инж. Ивайло Пенев, на колегите от катедра „Компютърни науки и технологии" ТУ-Варна, на моето семейство за оказаната помощ и подкрепа при разработване на настоящия дисертационен труд.

ABSTRACT

Dissertation Title: Research of machine learning methods for information encryption

Dimitar Georgiev Todorov

The main areas in cryptography are synchronous and asynchronous encryption. They are a means of protecting information and communications through a secret key. The main characteristic of each such system is its resistance to brute force attack - an attack with all possible variants of the key, which determines the great importance of its length. With the tremendous advances in computer technology in recent years and the increase in computing power, this resilience is declining. It can be determined that this trend definitely has an adverse effect on synchronous cryptography, which uses a secret key for a specific occasion. It is widely used due to its speed and easier implementation in various systems. This raises a problem with the stability of the synchronous cryptographic algorithm, the main solution of which over the years is to increase the length of the secret key [3,10,11,21,24,43,65,76,77,79,80,92,98,107,108,109].

The trend of increasing the computing power of public computer systems in recent years allows the realization of the possibilities of machine learning in different ways in cryptography. The biggest advantage of synchronous cryptographic algorithms is their speed and easier integration into different systems. These criteria determine the main motive for the work - providing an opportunity to increase the stability of synchronous cryptographic algorithms, using the capabilities of machine learning.

The object of research in the work is the stability of synchronous cryptographic algorithms.

The subject of research in this work is to increase the stability of synchronous cryptographic algorithms.

The main research hypothesis in the paper is the possibility to increase the stability of synchronous cryptographic algorithms using machine learning algorithms.

The work consists of 4 chapters:

1. **Machine learning and synchronous cryptography. Area of interaction. Organization and implementation of joint tasks**. This chapter discusses the resilience of synchronous cryptographic algorithms as an object and analysis of machine learning as a means of increasing the resilience of synchronous cryptographic algorithms.
2. **An approach to solving problems in synchronous cryptography through machine learning and an algorithm for placing data in a homogeneous environment**. This chapter of the dissertation presents a developed approach to the realization of cryptographic tasks through machine learning.
3. **Application of machine learning algorithms and algorithm for placing in a homogeneous environment for secret key recognition**. This chapter of the dissertation presents a practical application in a software product (application) of the selected algorithms for machine learning and the proposed algorithm for placing data in a homogeneous environment.

4. **Planning the experimental production. Conducting and analyzing experimental results**. This chapter presents the planning and conduct of experiments to evaluate the application of the proposed algorithm for placing data in a homogeneous environment and to evaluate its interaction with machine learning algorithms using 4 experiments with 26 trials.

The conducted experimental research and the obtained results support the research hypothesis of the research in the work and confirm that the proposed approach allows solving problems in synchronous cryptography using the algorithm for placing data in a homogeneous environment and classification with machine learning.