Mobil Cihaz Kriminalistikasında Yaddaş Seqmentlərinin Süni İntellekt Əsasında Təsnifatı

Rahib Ağababayev[1], Yadigar İmamverdiyev[2]
[1,2]Azərbaycan Texniki Universiteti, Bakı, Azərbaycan
[1]*rahib.agb@gmail.com*, [2]*yadigar.imamverdiyev@aztu.edu.az*

Xülasə— **Məqalədə mobil cihaz kriminalistikasında 4096 baytlıq yaddaş seqmentlərinin süni intellekt əsasında kriminalistik tip təsnifatı problemi araşdırılır. Anonimləşdirilmiş real mobil cihazdan əldə olunmuş 1692 nişanlanmış fayl əsasında 89177 seqmentdən ibarət dataset formalaşdırılmış və müxtəlif siniflər üzrə təsnifat aparılmışdır. Tədqiqatda bayt-statistik əlamətlərə əsaslanan klassik maşın öyrənməsi modelləri (Random Forest, XGBoost, SVM) xam bayt ardıcıllığı üzərində qurulmuş 1D-CNN modeli ilə müqayisə edilmişdir. Eksperimentlər göstərmişdir ki, XGBoost modeli ən yüksək nəticəni əldə etmiş, Random Forest modeli isə yüksək məhsuldarlıq nümayiş etdirməklə yanaşı, daha izah edilə bilən alternativ kimi çıxış etmişdir.**

Açar sözlər— mobil cihaz kriminalistikası; yaddaş seqmentlərinin təsnifatı; süni intellekt; maşın öyrənməsi; dərin öyrənmə; rəqəmsal sübutların analizi.

I. GİRİŞ

Müasir rəqəmsal mühitdə mobil cihazlar fotoşəkillər, videolar, səs yazıları, mesajlar, sənədlər, tətbiq artefaktları və istifadəçi davranışı ilə bağlı çoxsaylı rəqəmsal izləri özündə saxlayan əsas sübut mənbələrindən birinə çevrilmişdir. Bu səbəbdən mobil cihaz kriminalistikası yalnız verilənlərin çıxarılması mərhələsi ilə məhdudlaşmır; əsas problem çıxarılmış böyük həcmli və heterogen məlumatın qısa müddətdə, sistemli və etibarlı şəkildə təhlil edilməsidir. Süni intellekt və maşın öyrənməsi yanaşmalarının rəqəmsal kriminalistikaya inteqrasiyası məhz bu ehtiyacdan doğur. Burada məqsəd eksperti əvəz etmək deyil, ekspertin ilkin qərarverməsini dəstəkləyən intellektual mexanizm formalaşdırmaqdır. Bu baxımdan, rəqəmsal sübutların avtomatik aşkarlanması, qruplaşdırılması və prioritetləşdirilməsi müasir rəqəmsal kriminalistikanın ən aktual tədqiqat istiqamətlərindən biridir.

Mobil cihaz kriminalistikasında əsas çətinliklərdən biri böyük həcmli və heterogen rəqəmsal sübutların qısa müddətdə etibarlı şəkildə təhlil edilməsidir. Bu problem Azərbaycan Respublikasının informasiya təhlükəsizliyi və kibertəhlükəsizlik üzrə 2023–2027 Strategiyasında vurğulanan informasiya məkanının müasir təhdidlərdən qorunması, normativ bazanın təkmilləşdirilməsi və yeni texnoloji həllərin tətbiqi ehtiyacı ilə, eləcə də 2025–2028-ci illər üçün Süni İntellekt Strategiyasında qeyd edilən AI tədqiqatlarının təşviqi, tətbiq imkanlarının genişləndirilməsi və kadr potensialının inkişafı prioritetləri ilə uyğunluq təşkil edir.

Mövcud tədqiqatlar göstərir ki, süni intellekt və maşın öyrənməsi yanaşmaları rəqəmsal kriminalistikada fayl fraqmentlərinin təsnifatı, mətn əsaslı sübutların aşkarlanması, anomaliya deteksiyası və ilkin çeşidləmə (triage) üçün perspektivlidir; lakin real cihazlardan toplanmış etiketli verilənlərin məhdudluğu, modellərin izah edilə bilənliyi və nəticələrin real istintaq mühitinə ümumiləşdirilməsi hələ də açıq problemlər olaraq qalır.

Bu məqalədə həmin boşluğu qismən aradan qaldırmaq məqsədilə anonimləşdirilmiş real mobil cihazdan əldə olunmuş verilənlər əsasında 4096 baytlıq yaddaş seqmentlərinin kriminalistik tip təsnifatı həyata keçirilmiş, bayt-statistik əlamətlərə əsaslanan klassik maşın öyrənməsi modelləri ilə xam bayt ardıcıllığı üzərində qurulmuş 1D-CNN modeli vahid eksperimental protokol üzrə müqayisə edilmişdir. Tədqiqatın əsas töhfəsi seqment-səviyyəli real datasetin qurulması, maşın öyrənməsi (ing. Machine Learning, ML) və dərin öyrənmə yanaşmalarının eyni şərtlərdə qiymətləndirilməsi və mobil rəqəmsal kriminalistikada daha münasib model ailəsinin eksperimental əsaslandırılmasıdır. Məqalənin növbəti hissələrində əlaqəli işlər, məsələnin qoyuluşu, metodologiya, müqayisəli nəticələr və yekun nəticələr təqdim olunur.

II. ƏLAQƏLİ İŞLƏR

Rəqəmsal kriminalistikada süni intellekt və maşın öyrənməsi yanaşmaları fayl fraqmentlərinin təsnifatı, mətn əsaslı sübutların aşkarlanması, multimedia analizi, anomaliya deteksiyası və ilkin çeşidləmə kimi istiqamətlərdə geniş tətbiq olunur. Mövcud tədqiqatlar göstərir ki, nəzarətli maşın öyrənməsi və dərin öyrənmə modelləri böyük həcmli və heterogen rəqəmsal sübutların emalı üçün perspektivli nəticələr verir. Bununla belə, real cihazlardan toplanmış etiketli verilənlərin məhdudluğu, qiymətləndirmə protokollarının fərqliliyi və modellərin izah edilə bilənliyi hələ də sahənin əsas açıq problemləri olaraq qalır [1-4].

Bu məqalənin metodoloji çərçivəsi ilə yaxınlıq təşkil edən işlər arasında mətn analizi, anomaliyaların aşkarlanması, ansambl klassifikatorlar və klasterləşmə yanaşmaları xüsusi yer tutur. Terrorizmlə əlaqəli mətnlərin aşkarlanması üçün text mining əsaslı çoxmərhələli yanaşmada semantik yaxınlıq və hibrid təsnifatlandırmanın effektivliyi göstərilmişdir [5]. Təhlükəsizlik və böyük verilənlər mühitində çəkili klasterləşmə, çoxklassifikatorlu modellər və ansambl yanaşmaların anomaliyaların aşkarlanmasında səmərəli olduğu da göstərilmişdir [6-9]. Bu nəticələr heterogen verilənlərdə birdən çox model ailəsinin müqayisəsinin məqsədəuyğunluğunu əsaslandırır və təqdim olunan işdə Random Forest, XGBoost və SVM modellərinin paralel qiymətləndirilməsi üçün nəzəri baza

DOI: 10.25045/forensic.gov.2026.49

yaradır. Dərin öyrənmə ilə bağlı tədqiqatlar da göstərir ki, CNN və LSTM əsaslı modellər müəyyən problemlərdə yüksək nəticə verə bilir [10, 11]. Lakin bu üstünlük daha çox mətn və hadisə yönümlü verilənlər üzərində nümayiş etdirilmişdir. Mobil cihaz yaddaşından əldə olunmuş xam bayt ardıcıllıqları üçün bu nəticələrin birbaşa qorunub-qorunmaması ayrıca eksperimental yoxlama tələb edir. Bu çərçivədə, rəqəmsal kriminalistikada sübutların avtomatik təsnifatı üçün metadata əsaslı metodlar da təklif edilmişdir [12]; paralel olaraq, terrorizmlə əlaqəli mətnlərin filtrasiyası maşın öyrənməsi yanaşmaları ilə də tədqiq edilmişdir [13]. Buna görə bu məqalədə dərin öyrənmə yanaşması xam 4096 baytlıq ardıcıllıq üzərində qurulmuş sadə 1D-CNN modeli ilə qiymətləndirilmiş, klassik ML modelləri ilə eyni verilənlər toplusu və eyni fayl-mənbə səviyyəli bölgü protokolu üzrə müqayisə aparılmışdır.

Beləliklə, mövcud ədəbiyyat süni intellektin rəqəmsal kriminalistikada perspektivini təsdiqləsə də, real mobil cihazdan əldə olunmuş yaddaş seqmentlərinin kriminalistik tip təsnifatı üzrə klassik ML və dərin öyrənmə yanaşmalarının vahid eksperimental çərçivədə müqayisəsi hələ də məhdud işlənmiş istiqamət olaraq qalır. Təqdim olunan məqalənin yeniliyi məhz bu boşluğu qismən doldurmaqdan, yəni real verilənlər əsasında 4096 baytlıq yaddaş seqmentlərinin təsnifatını həyata keçirməkdən və hansı model ailəsinin daha uyğun olduğunu eksperimental olaraq göstərməkdən ibarətdir.

III. MƏSƏLƏNİN QOYULUŞU

Bu tədqiqatın məqsədi mobil cihazdan əldə olunmuş verilənlərin 4096 baytlıq yaddaş seqmentləri səviyyəsində avtomatik kriminalistik tip təsnifatını həyata keçirməkdir. Problem tam faylın analizi kimi deyil, kiçik sabit ölçülü seqmentlərin hansı sübut tipinə daha yaxın olduğunu müəyyən etmək kimi qoyulur. Belə yanaşma real rəqəmsal kriminalistika ssenarilərində bütöv və strukturlaşdırılmış fayllarla yanaşı, hissəvi, dağınıq və ya kontekstdən ayrılmış məlumat vahidlərinin təhlili ehtiyacını nəzərə alır.

Eksperimentdə anonimləşdirilmiş real mobil cihazdan əldə olunmuş verilənlər əsasında beş kriminalistik tip sinfi formalaşdırılmışdır: audio, sıxılmış-şifrlənmiş, şəkil, mətn və video. İlkin mərhələdə 1692 etiketlənmiş fayl seçilmiş, daha sonra həmin fayllar 4096 baytlıq seqmentlərə bölünərək ümumilikdə 89177 nümunədən ibarət dataset qurulmuşdur. Yeni fayl-mənbə səviyyəli bölgüdə 1183 təlim, 170 validasiya və 339 test faylı, seqment səviyyəsində isə 63 827 təlim, 8452 validasiya və 16898 test nümunəsi istifadə edilmişdir.

Bu məsələ iki fərqli model ailəsi çərçivəsində araşdırılır. Birinci yanaşmada seqmentlərdən çıxarılan bayt-statistik əlamətlər əsasında klassik maşın öyrənməsi modelləri tətbiq edilir. İkinci yanaşmada isə seqmentlərin xam bayt ardıcıllığı birbaşa giriş kimi istifadə olunur və bu əsasda 1D-CNN modeli qurulur. Beləliklə, tədqiqat yalnız təsnifat problemini deyil, eyni zamanda klassik ML və dərin öyrənmə yanaşmalarının vahid eksperimental çərçivədə müqayisəsini də əhatə edir (Şəkil 1).

Bu məqsədlə aşağıdakı tədqiqat sualları formalaşdırılmışdır:

RQ1. 4096 baytlıq yaddaş seqmentləri bayt-statistik əlamətlər əsasında kriminalistik tip siniflərinə yüksək dəqiqliklə ayrılırmı?

RQ2. Random Forest, XGBoost və SVM modelləri arasında mobil cihaz yaddaş seqmentlərinin çoxsinifli təsnifatında hansı model daha effektiv nəticə verir?

RQ3. Xam bayt ardıcıllığı üzərində qurulmuş sadə 1D-CNN modeli eyni problem üçün maşın öyrənməsi modelləri ilə rəqabət apara bilirmi?

RQ4. Mobil cihaz kriminalistikasında praktik tətbiq baxımından daha münasib yanaşma hansıdır: yüksək performanslı ansambl model, daha interpretasiya oluna bilən klassik model, yoxsa xam məlumat üzərində öyrənən dərin model?

Bu suallara uyğun olaraq iki əsas hipotez qəbul edilir. Birinci hipotez ondan ibarətdir ki, 4096 baytlıq yaddaş seqmentləri bayt-statistik baxımdan kifayət qədər fərqləndirici informasiya daşıdığı üçün klassik maşın öyrənməsi modelləri ilə effektiv şəkildə təsnif edilə bilər. İkinci hipotez isə odur ki, xam bayt ardıcıllığı üzərində öyrənən 1D-CNN nəzəri olaraq daha çevik nümayəndəlik yarada bilsə də, konkret bu dataset və seçilmiş sadə arxitektura şəraitində klassik modelləri mütləq şəkildə üstələməyə bilər.

Beləliklə, məsələnin qoyuluşu yalnız "seqment hansı sinfə aiddir?" sualına cavab verməkdən ibarət deyil. Eyni zamanda məqsəd mobil cihaz kriminalistikasında hansı model ailəsinin daha uyğun olduğunu eksperimental olaraq əsaslandırmaqdır.

IV. METODOLOGİYA

Bu bölmədə eksperimental mühit, datasetin hazırlanması, 4 kB seqmentləşdirmə, klassik ML və dərin öyrənmə modelləri, təlim strategiyası və qiymətləndirmə meyarları təqdim olunur.

A. *Eksperimental mühit*

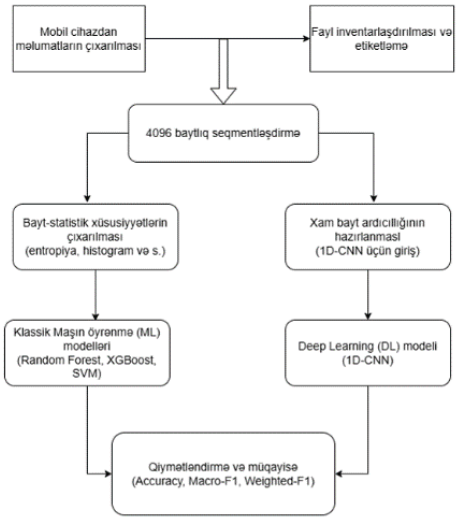
Tədqiqat anonimləşdirilmiş real mobil cihaz üzərində aparılmışdır. Eksperiment üçün cihazdan çıxarılan real fayl korpusu istifadə edilmiş, əməliyyat sisteminin fiziki yaddaş səhifələri deyil, kriminalistik analiz üçün uyğunlaşdırılmış 4096 baytlıq (4 kB) sabit yaddaş seqmentləri əsas vahid kimi qəbul olunmuşdur. Bu seçimin iki əsas səbəbi vardır: birincisi, sabit seqment ölçüsü təkrarolunan eksperimental protokol yaradır; ikincisi, həm bayt-statistik əlamətlərin çıxarılması, həm də xam bayt ardıcıllığı üzərində dərin modelin qurulması üçün yetərli informasiya daşıyır (Şəkil 1.).

Dərin öyrənmə modeli üçün eksperiment mühitində 1D-CNN modelinin yekun təlim və test mərhələləri CPU üzərində icra edilmişdir. Bu hal nəticələrin metodoloji bütövlüyünü pozmur, lakin gələcəkdə daha mürəkkəb arxitekturaların və daha geniş hiperparametr axtarışının GPU dəstəkli mühitdə sınaqdan keçirilməsinin məqsədəuyğun olduğunu göstərir.

B. *Datasetin hazırlanması*

İlkin inventarlaşdırma və filtrasiya mərhələsindən sonra eksperiment üçün 1692 etiketlənmiş fayl seçilmişdir. Fayl səviyyəsində sinif bölgüsü Cədvəl 1-də qeyd edilmişdir.

İlkin mərhələdə "no_extension", "tmp", "chck", "xlog", "dat" və digər yüksək səs-küylü fayl tipləri əsas eksperimentdən çıxarılmışdır. Məqsəd ilk mərhələdə nəzarət olunan və etibarlı etiketlənmiş dataset qurmaq olmuşdur.



Şəkil 1. Tədqiqatın ümumi blok sxemi

CƏDVƏL I. FAYL SƏVİYYƏSİNDƏ SİNİF BÖLGÜSÜ

Sinif	***Təsviri***	***Fayl sayı***
Audio	Səs yazıları	723
Sıxılmış/şifrlənmiş	Sıxılmış və ya şifrlənmiş	75
Şəkil	Şəkil	554
Mətn	Mətn faylı	20
Video	Video təsvir faylları	320
Cəmi		1692

C. *4096 baytlıq yaddaş seqmentlərinin formalaşdırılması*

Hər bir fayl 4096 baytlıq sabit hissələrə bölünmüşdür. Bu işdə "page" anlayışı NAND yaddaşının fiziki səhifəsi (native NAND physical page) kimi deyil, kriminalistik analiz üçün uyğunlaşdırılmış 4 kB bayt-seqment kimi istifadə edilmişdir. Seqmentləşdirmə nəticəsində ümumilikdə 89177 seqment alınmışdır: 3554 audio, 20925 şəkil, 62106 video, 1942 text_like və 650 compressed_encrypted seqment. Yeni bölgü (split) protokolunda seqment səviyyəli bölgü belə olmuşdur: 63827 təlim, 8452 validasiya və 16898 test nümunəsi.

$F = \{f_1, f_2, \ldots, f_n\}$ fayllar çoxluğu, $S = \{s_1, s_2, \ldots, s_n\}$ isə həmin fayllardan çıxarılan 4096 baytlıq seqmentlər çoxluğu olsun.

Kriminalistik tip sinifləri çoxluğu aşağıdakı kimi təyin edilir:

$$C = \{audio, compressed_{encrypted}, image, text_{like}, video\}$$

Məqsəd hər bir s_i seqmentini C çoxluğundakı uyğun siniflərdən birinə aid edən

$$g: S \to C$$

təsnifat funksiyasını qurmaqdır.

D. *Bayt-statistik əlamətlərin çıxarılması*

ML modelləri üçün hər 4096 baytlıq seqmentdən aşağıdakı əlamətlər çıxarılmışdır:

1. Şennon entropiyası
2. Sıfır bayt nisbəti
3. <ap edilə bilən simvolların payı
4. Baytların orta qiyməti
5. Standart sapma
6. 256 ölçülü normallaşdırılmış bayt histoqramı

Bu əlamətlər bayt-statistik profil yaratmaq üçün seçilmişdir. Entropiya yüksək sıxılmış və ya şifrlənmiş seqmentlərin ayrılması üçün, çap edilə bilən simvolların payı mətnəbənzər seqmentlərin aşkarlanması üçün, histoqram isə müxtəlif media və qeyri-media faylların bayt-paylanmasını modelləşdirmək üçün istifadə edilmişdir.

E. *Klassik maşın öyrənməsi (ML) modelləri*

Müqayisə üçün üç klassik model seçilmişdir:

- Random Forest
- XGBoost
- SVM-RBF

Random Forest yüksək nəticə ilə yanaşı əlamət əhəmiyyətlərinin çıxarılmasına imkan verdiyi üçün interpretasiya baxımından əlverişlidir. XGBoost çoxsinifli təsnifat problemlərində yüksək performans göstərən ansambl model kimi daxil edilmişdir. SVM isə klassik baza modeli kimi müqayisə üçün istifadə edilmişdir.

F. *Dərin öyrənmə modeli: 1D-CNN*

Dərin öyrənmə modelində seqmentlərin əvvəlcədən çıxarılmış əlamətləri deyil, onların xam 4096 baytlıq ardıcıllığı birbaşa giriş kimi istifadə edilmişdir. Bu məqsədlə sadə və nisbətən izah edilə bilən 1D-CNN arxitekturası qurulmuşdur.

Model üç ardıcıl konvolyusiya blokundan ibarətdir: Conv1D + BatchNorm + ReLU + MaxPool, Conv1D + BatchNorm + ReLU + MaxPool və Conv1D + BatchNorm + ReLU + MaxPool. Bundan sonra qlobal əlamətlərin sıxlaşdırılması üçün qlobal orta hovuzlama (Global Average Pooling), həddən artıq uyğunlaşmanın qarşısını almaq üçün "dropout" (təsadüfi neyron söndürülməsi), qərarvermə mərhələsi üçün tam əlaqəli qatlar və çoxsinifli təsnifat üçün "softmax" çıxış qatı istifadə edilmişdir. Bu arxitekturanın seçilməsində məqsəd ən mürəkkəb dərin öyrənmə modeli qurmaq deyil, klassik maşın öyrənməsi modelləri ilə müqayisə edilə bilən, təkrarolunan və metodoloji baxımdan şəffaf bir dərin öyrənmə bazası yaratmaq olmuşdur.

G. *Təlim strategiyası*

Bütün modellər üçün eyni prinsip saxlanılmışdır: təlim/validasiya/sınaq (train/validation/test) bölgüsü fayl-mənbə səviyyəsində aparılmışdır. Bu qərarın məqsədi məlumat sızmasının qarşısını almaqdır. Eyni fayldan çıxarılan seqmentlərin həm təlim, həm də test dəstinə düşməsi modelin süni yüksək nəticə verməsinə səbəb ola bilər; buna görə bölgü fayl səviyyəsində qurulmuşdur.

CNN üçün aşağıdakı təlim strategiyası tətbiq edilmişdir:

- optimallaşdırıcı (Adam)
- başlanğıc öyrənmə sürəti: 10
- maksimal epoxa sayı: 40
- erkən dayanma: 8 epoxa səbr limiti ilə
- model seçimi: ən yaxşı validation macro-F1
- sinif balanssızlığına qarşı: sinif çəkiləri
- dominant siniflərin həddən artıq təsirini azaltmaq üçün: seqment yanaşması

Burada istifadə olunan anlayışlar belə izah olunur:

Epoxa– modelin bütün train dəstini bir dəfə tam keçməsi.

Erkən dayandırma – validation nəticəsi yaxşılaşmadıqda təlimin vaxtından əvvəl dayandırılması.

Sinif çəkiləri – az təmsil olunan siniflərin itki funksiyasında daha böyük çəkidə nəzərə alınması.

Ən yaxşı validasiya dəstində Macro-F1 – ən yaxşı modelin yalnız train itkiyə görə deyil, validasiya dəstində siniflərarası balanslı məhsuldarlığa görə seçilməsi.

H. *Qiymətləndirmə metrikaları*

Tədqiqatda aşağıdakı metrikalardan istifadə edilmişdir:

1. Accuracy Ümumi düzgün proqnozların bütün nümunələrə nisbətidir:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

Çoxsinifli halda bu göstərici ümumi düzgün təsnif edilmiş nümunələrin payını göstərir.

2. Precision – Müəyyən sinif üzrə modelin "bu sinfə aiddir" dediyi nümunələrin nə qədərinin həqiqətən həmin sinfə aid olduğunu göstərir:

$$Precision_c = \frac{TP_C}{TP_C+FP_C} \qquad (2)$$

3. Recall – Müəyyən sinfə aid olan real nümunələrin nə qədərinin model tərəfindən düzgün aşkar edildiyini göstərir:

$$Recall_C = \frac{TP_C}{TP_C+FN_C} \qquad (3)$$

4. F1-score – Precision və recall-un harmonik ortasıdır:

$$F1_C = 2 \cdot \frac{Precision_C \cdot Recall_C}{Precision_C+Recall_C} \qquad (4)$$

5. Macro-F1 – Bütün siniflər üzrə F1-score-ların sadə ədədi ortasıdır:

$$Macro - F1 = \frac{1}{|C|}\sum_{c\epsilon C} F1_c \qquad (5)$$

Bu göstərici hər sinfə bərabər çəki verir və balanssız siniflər üçün əsas metrikadır.

6. Weighted-F1 – Hər sinfin F1-score göstəricisinin həmin sinfin nümunə sayına görə çəkili ortasıdır:

$$Weighted - F1 = \sum_{c\epsilon C} \frac{n_c}{N} F1_c \qquad (6)$$

7. Qarışıqlıq matrisi (Confusion Matrix) – Modelin hansı sinfi hansı siniflə daha çox qarışdırdığını göstərən cədvəldir. Bu, xüsusilə şəkil-video və text_like-compressed_encrypted kimi yaxın strukturlu siniflər arasındakı qarışıqlığı təhlil etməyə imkan verir.

Siniflər balanslı olmadığından bu tədqiqatda əsas vurğu macro-F1 göstəricisinə verilmişdir.

V. TƏDQİQAT NƏTİCƏLƏRİ VƏ MÜZAKİRƏ

Eksperimentlər vahid fayl-mənbə səviyyəli train/validation/test protokolu əsasında aparılmışdır. Yeni bölgü Cədvəl 2-də qeyd edilmişdir. Bu bölgü həm klassik ML, həm də dərin öyrənmə modellərinin eyni şərtlər altında müqayisəsinə imkan vermişdir.

CƏDVƏL II. TRAIN/VALIDATION/TEST ÜZRƏ SEQMENT BÖLGÜSÜ

Bölgü (Split)	***Təlim***	***Validasiya***	***Test***	***Cəmi***
Audio	2549	361	644	3554
Sıxılmış/şifrlənmiş	350	79	221	650
Şəkil	15357	1748	3820	20925
Mətn	1640	30	272	1942
Video	43931	6234	11941	62106
Ümumi seqment sayı	63827	8452	16898	89177

Klassik maşın təlim modelləri ilə 1D-CNN modelinin müqayisəli nəticələri **Cədvəl 3**-də təqdim olunmuşdur.

CƏDVƏL III. MODELLƏRIN MÜQAYISƏLI NƏTICƏLƏRI

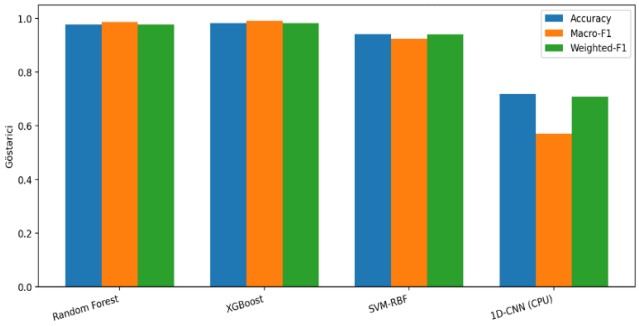
Model	***Accuracy***	***Macro-F1***	***Weighted-F1***
Random Forest	0.977	0.9858	0.9766
XGBoost	0.9826	0.9896	0.9824
SVM-RBF	0.9412	0.9246	0.9402
1D-CNN	0.7183	0.5701	0.7090

Cədvəl 3-dəki nəticələr ilk növbədə onu göstərir ki, bayt-statistik əlamətlərə əsaslanan ML yanaşmaları bu problem üçün yüksək diskriminativ gücə malikdir. Xüsusilə macro-F1 göstəricisinin həm XGBoost, həm də Random Forest üçün 0.98-dən yüksək olması mühüm əhəmiyyət daşıyır. Datasetdə video sinfi dominant olduğundan yalnız accuracy və weighted-F1 göstəricilərinə əsaslanmaq kifayət etmir. Macro-F1 isə bütün siniflərə bərabər çəki verdiyi üçün modelin dominant sinifdən kənarda qalan sinifləri də nə dərəcədə düzgün ayırdığını göstərir. Bu baxımdan həm XGBoost, həm də Random Forest modellərinin siniflərarası ayırdetmə qabiliyyətinin yüksək olduğu görünür.

Siniflər üzrə təhlil modellərin güclü tərəflərini daha aydın göstərir. Audio və compressed_encrypted sinifləri test dəstində həm Random Forest, həm də XGBoost tərəfindən faktiki olaraq

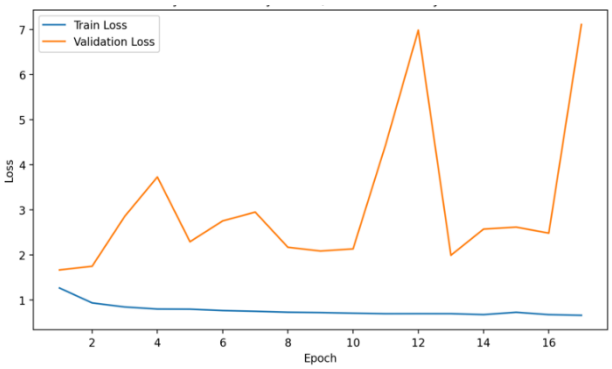
ideal şəkildə təsnif edilmişdir. Text_like sinfi də xüsusilə XGBoost üçün problemsiz ayrılmışdır. Ən çətin sinif şəkil (image) olmuşdur: Random Forest modelində recall 0.9018, XGBoost modelində isə 0.9283 olmuşdur. Bu hal şəkil və video seqmentləri arasında bayt-statistik baxımdan müəyyən yaxınlığın qalması ilə izah edilə bilər. Bununla belə, hər iki model həmin sinifdə də yüksək F1-score saxlamışdır. Modellərin accuracy, macro-F1 və weighted-F1 göstəriciləri üzrə müqayisəsi Şəkil 2-də verilmişdir.

Validation nəticələrində xüsusilə compressed_encrypted sinfi üzrə aşağı recall müşahidə edilmişdir. Bu vəziyyət modelin ümumi zəifliyindən çox, validation dəstində həmin sinfin çox az nümunə ilə təmsil olunması ilə əlaqəlidir. Test dəstində eyni sinif üzrə modellərin tam düzgün nəticə verməsi bu qənaəti gücləndirir. Başqa sözlə, azsaylı siniflər üçün validation nəticələri yüksək variasiya göstərə bilər; buna görə yekun şərh əsasən test dəsti üzərində qurulmalıdır.



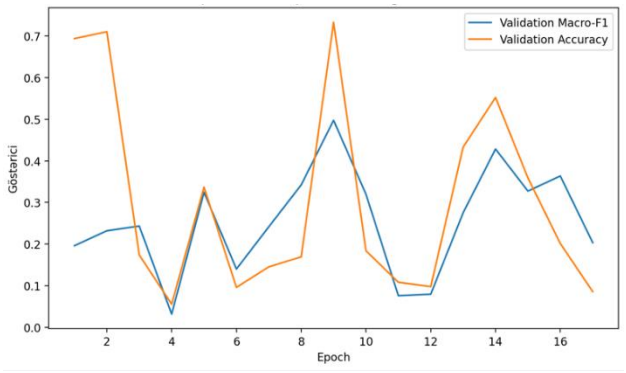
Şəkil 2. Modellərin test nəticələrinin müqayisəsi

Dərin öyrənmə yanaşmasında qurulmuş sadə 1D-CNN modeli həm validation, həm də test mərhələsində modellərdən nəzərəçarpacaq dərəcədə zəif olmuşdur. Ən yaxşı validation macro-F1 0.4973, ən yaxşı epoxa isə 9-cu epoxa olmuş, model 17-ci epoxada erkən dayandırılmışdır. Test mərhələsində CNN audio sinfində yüksək nəticə göstərsə də, compressed_encrypted və şəkil siniflərində zəif performans nümayiş etdirmişdir. Bu, xam bayt ardıcıllığı üzərində qurulmuş sadə 1D-CNN modelinin bu datasetdə sabit və kifayət qədər diskriminativ nümayəndəlik yarada bilmədiyini göstərir. CNN-in loss əyriləri Şəkil 3-də, validasiya göstəriciləri isə Şəkil 4-də göstərilmişdir.



Şəkil 3. CNN üçün train/validation loss əyriləri

CNN modelinin ML modellərindən geri qalmasının bir neçə texniki səbəbi vardır. Birincisi, Random Forest, XGBoost və SVM əvvəlcədən çıxarılmış bayt-statistik əlamətlər üzərində işlədiyi halda, CNN bütün fərqləndirici siqnalları xam 4096 baytlıq ardıcıllıqdan özü öyrənməlidir. İkincisi, seqment sayı böyük olsa da, müxtəliflik məhdud sayda mənbə fayldan gəldiyinə görə dərin model üçün effektiv müstəqil nümunə sayı nəzəri göründüyü qədər böyük deyil. Üçüncüsü, sinif balanssızlığı, xüsusilə video sinfinin dominantlığı, CNN-in az təmsil olunan siniflər üçün sabit qərar sərhədi formalaşdırmasını çətinləşdirmişdir. Dördüncüsü, bu işdə qəsdən sadə və interpretasiya baxımından nisbətən şəffaf 1D-CNN arxitekturası seçilmişdir; buna görə daha mürəkkəb dərin arxitekturaların potensial üstünlüyü ayrıca məsələ olaraq qalır.



Şəkil 4. CNN üçün validasiya göstəriciləri

Bu nəticə akademik baxımdan dərin öyrənmənin istənilən halda ML-dən üstün olduğunu göstərmək deyil, konkret problem və konkret verilənlər çoxluğu üçün hansı yanaşmanın daha uyğun olduğunu eksperimental olaraq müəyyən etməy' imkan verir. Bu işdə bütün modellər eyni split və eyni qiymətləndirmə protokolu üzrə sınaqdan keçirilmişdir. Buna görə də alınmış nəticə elmi baxımdan mənfi nəticə kimi qiymətləndirilməməlidir, əksinə, müqayisəli və metodoloji baxımdan dəyərli nəticədir: mobil cihaz yaddaş seqmentlərinin kriminalistik tip təsnifatı problemində model mürəkkəbliyinin artırılması avtomatik olaraq daha yüksək performans verməmişdir.

Praktik baxımdan rəqəmsal kriminalistikada model seçimi yalnız accuracy ilə deyil, həm də sabitlik, izah edilə bilmə, audit edilə bilmə və ekspert qarşısında müdafiə oluna bilmə kimi meyarlarla müəyyən edilir. Bu baxımdan XGBoost performans lideri, Random Forest isə praktik tətbiq və izah edilə bilənlik baxımından daha münasib model kimi qiymətləndirilə bilər. Beləliklə, bu tədqiqatın əsas nəticəsi ondan ibarətdir ki, mobil cihaz kriminalistikasında 4 kB yaddaş seqmentlərinin kriminalistik tip təsnifatı üçün hazırkı verilənlər toplusu və eksperiment quruluşunda ML modelləri, xüsusilə XGBoost və Random Forest, sadə 1D-CNN modelindən daha yüksək və daha sabit nəticə nümayiş etdirmişdir.

NƏTİCƏ

Məqalədə mobil cihaz kriminalistikasında 4096 baytlıq yaddaş seqmentlərinin süni intellekt əsasında kriminalistik tip təsnifatı problemi tədqiq edilmişdir. Anonimləşdirilmiş real mobil cihazdan əldə olunmuş 1692 etiketlənmiş fayl əsasında 89177 seqmentdən ibarət verilənlər toplusu qurulmuş, audio, sıxılmış-şifrlənmiş, şəkil, mətn və video sinifləri üzrə həm

klassik ML, həm də dərin öyrənmə yanaşmaları müqayisə edilmişdir. Eyni fayl-mənbə səviyyəli təlim/validasiya/sınaq protokolu üzrə aparılmış eksperimentlərdə XGBoost ən yüksək nəticəni göstərmişdir (accuracy 0.9826, macro-F1 0.9896), Random Forest isə ona çox yaxın performansla yanaşı daha interpretasiya oluna bilən model kimi çıxış etmişdir (accuracy 0.9770, macro-F1 0.9858). 1D-CNN modeli isə bu datasetdə klassik yanaşmaları üstələyə bilməmişdir (accuracy 0.7183, macro-F1 0.5702).

Alınmış nəticələr göstərir ki, mobil cihaz yaddaş seqmentlərinin kriminalistik tip təsnifatı problemində bayt-statistik əlamətlərə əsaslanan ML modelləri, xüsusilə XGBoost və Random Forest, hazırkı quruluşda xam bayt ardıcıllığı üzərində öyrənən sadə 1D-CNN modelindən daha yüksək və daha sabit nəticə verir. Təklif olunan yanaşma böyük həcmli mobil məlumatın ilkin qruplaşdırılması, sübutların filtrasiya olunması və ekspert diqqətinin prioritet seqmentlərə yönəldilməsi üçün praktik baza yarada bilər.

Gələcək işlərdə datasetin çoxcihazlı mühitdə genişləndirilməsi, unknown/other sinfinin daxil edilməsi, deleted artifacts ssenarisinin ayrıca qiymətləndirilməsi, daha mürəkkəb CNN və Transformer əsaslı dərin modellərin sınağı, eləcə də interpretasiya analizinin genişləndirilməsi məqsədəuyğun görünür. Bu mənada təqdim olunan iş mobil cihaz kriminalistikasında süni intellekt əsaslı seqment-səviyyəli analiz üçün ilkin, lakin möhkəm metodoloji baza formalaşdırır.

ƏDƏBİYYAT

[1] R. M. Alıquliyev, G. Y. Niftəliyeva, "Text mining metodlarının köməyilə e-dövlətdə terrorizmlə əlaqəli məqalələrin aşkarlanması," İnformasiya Texnologiyaları Problemləri, vol. 6, no. 2, pp. 41–52, 2015.

[2] R. M. Alguliyev, R. M. Aliguliyev, Y. N. Imamverdiyev, and L. V. Sukhostat, "Weighted clustering for anomaly detection in Big Data," Statistics, Optimization and Information Computing, vol. 6, no. 2, pp. 178–188, 2018, doi: 10.19139/soic.v6i2.404.

[3] R. M. Alıquliyev, G. Y. Niftəliyeva, "Text mining metodlarının köməyilə e-dövlətdə terrorizmlə əlaqəli məqalələrin aşkarlanması," İnformasiya Texnologiyaları Problemləri, vol. 6, no. 2, pp. 41–52, 2015.

[4] R. M. Alguliyev, R. M. Aliguliyev, Y. N. Imamverdiyev, and L. V. Sukhostat, "Weighted clustering for anomaly detection in Big Data," Statistics, Optimization and Information Computing, vol. 6, no. 2, pp. 178–188, 2018, doi: 10.19139/soic.v6i2.404.

[5] R. M. Alguliyev, R. M. Aliguliyev, Y. N. Imamverdiyev, and L. V. Sukhostat, "An improved ensemble approach for DoS attacks detection," Radio Electronics, Computer Science, Control, no. 2, pp. 73–82, 2018, doi: 10.15588/1607-3274-2018-2-8.

[6] R. M. Alguliyev and M. Sh. Hajirahimova, "Classification ensemble based anomaly detection in network traffic," Review of Computer Engineering Research, vol. 6, no. 1, pp. 12–23, 2019, doi: 10.18488/journal.76.2019.61.12.23.

[7] R. M. Alguliyev, R. M. Aliguliyev and F. J. Abdullayeva, "The improved LSTM and CNN models for DDoS attacks prediction in social media," International Journal of Cyber Warfare and Terrorism, vol. 9, no. 1, pp. 1–18, 2019, doi: 10.4018/IJCWT.2019010101.

[8] X. Du, C. Hargreaves, J. Sheppard, F. Anda, A. P. Sayakkara, N.-A. Le-Khac and M. Scanlon, "SoK: Exploring the state of the art and the future potential of artificial intelligence in digital forensic investigation," in Proc. 15th Int. Conf. Availability, Reliability and Security (ARES), 2020, pp. 46:1–46:10.

[9] S. Fitzgerald, G. Mathews, C. Morris and O. Zhulyn, "Using NLP techniques for file fragment classification," Digital Investigation, vol. 9, suppl. 1, pp. S44–S49, 2012.

[10] M. Bhatt, A. Mishra, M. W. U. Kabir, S. E. Blake-Gatto, R. Rajendra, M. T. Hoque, and I. Ahmed, "Hierarchy-Based File Fragment Classification," Machine Learning and Knowledge Extraction, vol. 2, no. 3, pp. 216–232, 2020, doi: 10.3390/make2030012.

[11] X. Du, N.-A. Le-Khac, and M. Scanlon, "Deep learning for digital forensics: A survey," Forensic Science International: Digital Investigation, 2020.

[12] "Azərbaycan Respublikasının informasiya təhlükəsizliyi və kibertəhlükəsizliyə dair 2023–2027-ci illər üçün Strategiyası," Azərbaycan Respublikası Prezidentinin rəsmi sənədi, 2023.

[13] "Azərbaycan Respublikasının 2025–2028-ci illər üçün süni intellekt Strategiyası," Azərbaycan Respublikası Prezidentinin rəsmi sənədi, 2025.

[14] X. Du and M. Scanlon, "Methodology for the automated metadata-based classification of incriminating digital forensic artefacts," in Proc. 14th Int. Conf. Availability, Reliability and Security (ARES), 2019.

[15] R. M. Alguliyev, R. M. Aliguliyev, and G. Y. Niftaliyeva, "Filtration of terrorism-related texts in the e-government environment," International Journal of Cyber Warfare and Terrorism, vol. 8, no. 4, pp. 35–48, 2018.

Artificial Intelligence-Based Classification of Memory Segments in Mobile Device Forensics

Rahib Aghababayev[1], Yadigar Imamverdiyev[2]

[1,2]Azerbaijan Technical University, Baku, Azerbaijan

[1]*rahib.agb@gmail.com*, [2]*yadigar.imamverdiyev@aztu.edu.az*

Abstract- This paper investigates artificial intelligence-based forensic type classification of 4096-byte memory segments in mobile device forensics. A dataset of 89,177 segments was constructed from 1,692 labeled files acquired from an anonymized real mobile device. The segments were classified into five categories: audio, compressed_encrypted, image, text_like, and video. Classical machine learning models based on byte-statistical features, namely Random Forest, XGBoost, and SVM, were compared with a 1D-CNN trained on raw byte sequences. The results show that classical ML approaches, particularly XGBoost and Random Forest, provide more accurate and stable performance than the evaluated deep learning model.

Keywords- mobile device forensics; memory segment classification; artificial intelligence; machine learning; deep learning; digital evidence analysis.