Deepfake Əsaslı İnteraktiv Avatarların Təhlükəsizlik Təhlili və Biometrik Autentifikasiya Sistemləri üçün Riskləri

Ziyafət Əmirov[1], Məcid Qurbanov[2], Abbas Hüseynzadə[3]
[1,2,3]Dövlət Təhlükəsizliyi Xidmətinin Heydər Əliyev adına Akademiyası, Bakı, Azərbaycan
[1]*zamirov@inbox.ru*, [2]*majidgurbanov@gmail.com*

***Xülasə*— Generativ süni intellekt sahəsində əldə olunan son nailiyyətlər deepfake texnologiyalarının funksional imkanlarını əhəmiyyətli dərəcədə genişləndirərək süni intellekt əsaslı interaktiv sistemlərin formalaşmasına şərait yaratmışdır. Bu sistemlər vizual, audio və davranış xüsusiyyətlərini inteqrasiya edərək real-vaxt rejimində yüksək realizm səviyyəsinə malik imitasiya imkanları təqdim edir və biometrik autentifikasiya mexanizmləri üçün ciddi təhlükəsizlik riskləri yaradır. Məqalədə deepfake əsaslı interaktiv sistemlərin arxitekturası, hücum modelləri və biometrik identifikasiya alqoritmlərinə təsiri sistemli şəkildə təhlil edilmişdir. Həmçinin mövcud aşkarlama metodları və multimodal müdafiə yanaşmaları qiymətləndirilir. Nəticələr adaptiv və çoxqatlı təhlükəsizlik strategiyalarının zəruriliyini göstərir.**

Açar sözlər*— *dərin saxtalaşdırma (Deepfake); rəqəmsal avatar; generativ rəqib şəbəkəsi; variasiya avtoenkoderi.

I. GİRİŞ

Müasir identifikasiya və autentifikasiya sistemləri biometrik məlumatlara əsaslanır və xüsusilə sifətin və səsin tanınması texnologiyaları geniş tətbiq olunur. Bu yanaşma autentifikasiya prosesinin dəqiqliyini və istifadə rahatlığını artırsa da, dərin saxtalaşdırma (Deepfake) texnologiyalarının sürətli inkişafı yeni təhlükəsizlik risklərinin yaranmasına səbəb olmuşdur [1]. Bu risklərin əsas səbəbi Deepfake sistemlərində istifadə olunan generativ modellərin sürətlə təkmilləşməsi və daha realistik sintetik məlumatlar istehsal edə bilməsidir [2].

Bu texnologiyalar müxtəlif generativ modellərin inteqrasiyası əsasında formalaşır. Generativ Rəqib Şəbəkələr (Generative Adversarial Network - GAN) generator və diskriminator arasında qarşılıqlı rəqabət prinsipi əsasında yüksək keyfiyyətli görüntülərin yaradılmasını təmin edir [3]. Bu sahədə xüsusilə StyleGAN və StyleGAN2 kimi modellər realistik insan üzlərinin sintezində geniş istifadə olunur [4]. Digər tərəfdən, Variasiyalı Avtokodlayıcılar (Variational Autoencoder - VAE) insanın emosional və mimik xüsusiyyətlərinin latent məkan daxilində modelləşdirilməsinə imkan yaradır [5].

Son dövrlərdə diffuziya modelləri, xüsusilə Stable Diffusion kimi sistemlər vizual məlumatların daha stabil və detallı şəkildə generasiya olunmasını təmin edərək sintetik məzmunun realizm səviyyəsini əhəmiyyətli dərəcədə artırmışdır. Bu modellərin sintezi nəticəsində artıq yalnız statik saxta görüntülər deyil, eyni zamanda insan davranışlarını, mimikalarını və danışıq üslubunu real-vaxtda modelləşdirən interaktiv avatarlar və sintetik rəqəmsal kimliklər formalaşmışdır [6].

Bu cür sistemlərdə vizual, audio və dil modellərinin inteqrasiyası mühüm rol oynayır və multimodal süni intellekt yanaşmaları bu sahədə əsas texnoloji baza hesab olunur [7]. Bu kontekstdə böyük dil modelləri sintetik avatarların istifadəçi ilə daha təbii, kontekstual və dinamik ünsiyyət qurmasına imkan verir.

Belə tam funksional sistemlər real insan davranışını yüksək dəqiqliklə imitasiya edə bildiyi üçün mövcud biometrik autentifikasiya mexanizmlərini aldatmaq potensialına malikdir [8]. Bu texnologiyalar eyni zamanda sosial mühəndislik hücumlarının effektivliyini əhəmiyyətli dərəcədə artırır . Bəzi hallarda interaktiv avatarların yaradılması üçün tələb olunan vizual və səs məlumatları Deepfake texnologiyaları vasitəsilə sintetik şəkildə generasiya oluna bilər ki, bu da real şəxslərin razılığı olmadan onların simasında sistemlərin yaradılmasına imkan verir.

II. ƏLAQƏLİ İŞLƏR

Deepfake texnologiyaları ilə bağlı mövcud elmi tədqiqatlar sintetik məzmunun yaradılması, bu məzmunun aşkarlanması və müxtəlif tətbiq sahələrinin araşdırılması kimi istiqamətləri əhatə edir [2].

Generasiya sahəsində ilkin yanaşmalar əsasən GAN və VAE modellərinə əsaslanmışdır; burada GAN yanaşması generator və diskriminator arasında rəqabət mexanizmi vasitəsilə realistik görüntülər yaradır, VAE isə məlumatı latent məkan daxilində sıxlaşdıraraq əsas xüsusiyyətlərin qorunması ilə yenidən qurur [3], [5].

Bu istiqamətdə inkişaf etdirilmiş generativ modellər, xüsusilə GAN əsaslı yanaşmalar, üz görüntülərinin yüksək realizm səviyyəsində sintezində geniş tətbiq olunmuşdur [4].

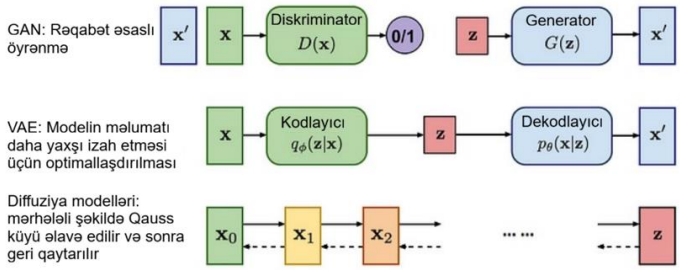
Son illərdə generativ süni intellekt sahəsində diffuziya əsaslı yanaşmalar ön plana çıxmışdır. Bu üsul mərhələli şəkildə küyün təmizlənməsi prinsipi ilə işləyərək daha sabit və inandırıcı nəticələr əldə etməyə imkan verir [9]. Xüsusilə latent diffuziya modelləri bu yanaşmanın praktik tətbiqini genişləndirmiş və yüksək keyfiyyətli görüntü sintezini mümkün etmişdir [10].

Aparılan tədqiqatlar göstərir ki, diffuziya əsaslı metodlar bir

DOI: 10.25045/forensic.gov.2026.04

çox hallarda əvvəlki GAN yanaşmalarını həm stabillik, həm də vizual realizm baxımından üstələyir [11].

Generativ modellərin - GAN, VAE və diffuziya əsaslı yanaşmaların - ümumi iş prinsipləri və fərqli mexanizmləri Şəkil 1-də müqayisəli şəkildə təqdim olunur.



Şəkil 1. Generativ modellərin müqayisəsi: GAN, VAE və diffusion əsaslı yanaşmalar.

Şəkil 1-də göstərildiyi kimi, GAN modelində həm real məlumat (X), həm də generator tərəfindən yaradılan sintetik məlumat (X') diskriminatora (D(x)) təqdim olunur və bu model məlumatın real və ya saxta olduğunu (0/1) müəyyən edir, generator isə latent dəyişəndən (z) istifadə edərək getdikcə daha realistik nəticələr istehsal etməyə çalışır. VAE modelində giriş məlumatı (X) kodlayıcı vasitəsilə latent məkan daxilində (z) ehtimal əsaslı şəkildə təmsil olunur və dekodlayıcı vasitəsilə yenidən qurularaq (X') bərpa edilir. Diffuziya modellərində isə başlanğıc məlumat (X_0) üzərinə mərhələli şəkildə küy əlavə olunur (X_1, X_2, …) və daha sonra bu proses tərsinə icra edilərək ilkin məlumat bərpa edilir. Bu yanaşmalar göstərir ki, generativ modellər fərqli mexanizmlərlə işləsə də, hamısı yüksək realizm səviyyəsinə malik sintetik məlumatların yaradılmasına yönəlib.

Deepfake aşkarlanması sahəsində isə müxtəlif metodlar təklif edilmişdir. Bu istiqamətdə XceptionNet və EfficientNet kimi dərin öyrənmə yanaşmaları vizual uyğunsuzluqların və artefaktların müəyyən edilməsi üçün geniş istifadə olunur [12].

Bəzi metodlar temporal analiz vasitəsilə sifət hərəkətlərini və mimikanı qiymətləndirir, digərləri isə bioloji göstəricilərə, məsələn göz qırpma tezliyi və üz simmetriyasına əsaslanır [13]. Lakin generativ modellərin inkişafı bu xüsusiyyətlərin də imitasiya olunmasını mümkün edir və nəticədə mövcud aşkarlama üsullarının effektivliyi azalır [1].

Bununla yanaşı, bu texnologiyaların sosial və hüquqi təsirləri də geniş şəkildə araşdırılmışdır. Araşdırmalar göstərir ki, Deepfake texnologiyaları dezinformasiya yaymaq, reputasiyaya zərər vurmaq və sosial mühəndislik hücumlarını həyata keçirmək üçün istifadə oluna bilər. Xüsusilə siyasi proseslər, ictimai rəyin formalaşması və maliyyə sistemləri bu təsirlərə daha həssas sahələr kimi qeyd olunur [14].

Buna baxmayaraq, interaktiv avatarların təhlükəsizlik baxımından analizi hələ də kifayət qədər inkişaf etməmişdir. Mövcud tədqiqatların əksəriyyəti statik Deepfake materiallara fokuslanır, halbuki interaktiv sistemlər real-vaxt rejimində cavab verə bilir, istifadəçi ilə ünsiyyət qurur və daha mürəkkəb davranış modelləri nümayiş etdirir. Bu xüsusiyyətlər onları daha təhlükəli edir və klassik yanaşmalardan fərqli olaraq daha geniş risklər yaradır.

III. İNTERAKTİV AVATARLARIN ARXİTEKTURASI VƏ FUNKSİONAL XÜSUSİYYƏTLƏRİ

İnteraktiv avatarların texniki strukturu yalnız müxtəlif komponentlərin birləşdirilməsi ilə məhdudlaşmır, eyni zamanda bu komponentlərin bir-biri ilə uyğun və koordinasiyalı şəkildə işləməsini tələb edir [7]. Bu sistemlər real-vaxt rejimində işlədiyi üçün məlumatların sürətli emalı və modellər arasında düzgün əlaqə çox vacibdir. Müasir arxitekturalarda adətən bir neçə model paralel şəkildə fəaliyyət göstərir və hər biri konkret funksiyanı yerinə yetirir. Bu modellərin birgə işləməsi nəticəsində isə vahid və uyğun davranış göstərən sintetik avatar formalaşır.

Vizual komponent yalnız statik sifət şəkillərinin yaradılması ilə kifayətlənmir. Bu model zaman daxilində üz hərəkətlərini də izləyir və modelləşdirir. Yəni video ardıcıllığında hər bir kadr əvvəlki ilə uyğun olur və bu da görüntünün daha real görünməsini təmin edir [6], [15]. Bu cür ardıcıllıq Deepfake videolarının aşkar olunmasını çətinləşdirir. Bundan əlavə, müasir sistemlər işıqlandırma, kölgə və perspektiv kimi detalları da nəzərə alır. Nəticədə yaradılan görüntülər daha təbii və inandırıcı olur.

Səs sintez sistemləri də son illərdə sürətlə inkişaf etmişdir. Xüsusilə transformer əsaslı modellərin tətbiqi bu sahədə keyfiyyət artımına səbəb olmuşdur [16]. Bu sistemlər yalnız sözlərin düzgün tələffüzünü təmin etmir, həm də danışığın emosional tərəfini modelləşdirə bilir. Məsələn, eyni cümlə fərqli emosiyalarla səsləndirilə bilər və bu, sintetik səsi daha inandırıcı edir. Bu səbəbdən belə sistemlər yalnız texniki baxımdan deyil, həm də psixoloji təsir baxımından real insan səsinə yaxınlaşır.

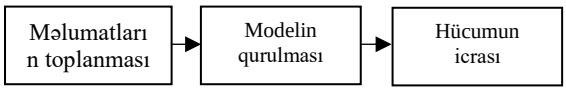
Böyük dil modellərinin bu sistemlərə inteqrasiyası isə onların funksionallığını daha da artırır. Bu modellər istifadəçinin dediklərini anlayaraq uyğun cavablar yarada bilir və uzunmüddətli dialoqları yadda saxlaya bilir [17], [18]. Bu xüsusiyyət sayəsində interaktiv avatarlar istifadəçi ilə təbii ünsiyyət qura bilir və real insan kimi davranır. Məhz bu cür davranış imkanları interaktiv avatarların sosial mühəndislik hücumlarında daha effektiv istifadə olunmasına şərait yaradır, çünki sistem qarşı tərəfdə inam hissi formalaşdıra bilir.

IV. TƏHLÜKƏSİZLİK TƏHLİLİ VƏ HÜCUM MODELLƏRİ

İnteraktiv avatarlara əsaslanan hücumlar müasir biometrik autentifikasiya sistemləri üçün ciddi təhlükəsizlik riskləri yaradır. Bu tip hücumların effektivliyi yalnız generativ modellərin keyfiyyəti ilə deyil, həm də hədəf sistemin arxitekturası, istifadə olunan biometrik alqoritmlər və autentifikasiya mexanizmlərinin mürəkkəblik səviyyəsi ilə müəyyən olunur. Xüsusilə üz və səs əsaslı identifikasiya sistemləri bu cür hücumlara qarşı daha həssasdır, çünki onlar əsasən statistik və xüsusiyyət əsaslı müqayisələrə əsaslanır və yüksək realizm səviyyəsinə malik sintetik məlumatlarla aldadıla bilər [8].

Deepfake əsaslı interaktiv avatar hücumlarını sistemli şəkildə təhlil etmək üçün bu prosesi mərhələli model əsasında nəzərdən keçirmək məqsədəuyğundur. Bu yanaşmaya görə hücum prosesi üç əsas mərhələdən ibarətdir: məlumatların toplanması, modelin qurulması və hücumun icrası. Təklif olunan

mərhələli hücum modelinin ümumi arxitekturası Şəkil 2-də təqdim olunur.



Şəkil 2. Deepfake əsaslı interaktiv avatar hücum modelinin mərhələli arxitekturası

1. Bu mərhələdə hücum edən tərəf hədəf istifadəçi haqqında multimodal məlumat toplayır. Buraya yalnız statik şəkillər deyil, müxtəlif bucaqlardan çəkilmiş videolar, fərqli emosional vəziyyətlərdə səs yazıları və davranış xüsusiyyətləri daxildir. Sosial şəbəkələr, video platformalar və açıq mənbələr bu məlumatların əsas mənbəyidir.

Toplanmış məlumatlar daha sonra "synthetic identity profile" kimi strukturlaşdırılır. Bu profil istifadəçinin yalnız fiziki xüsusiyyətlərini deyil, həm də danışıq tempi, reaksiyaları və ünsiyyət üslubunu əhatə edir. Bu isə hücumun sosial mühəndislik baxımından effektivliyini artırır. Əksər interaktiv avatar platformaları istifadəçinin video və səs məlumatlarına əsaslansa da, Deepfake texnologiyaları bu məlumatların süni şəkildə generasiya olunmasına imkan verir və nəticədə real şəxsin razılığı olmadan onun simasında interaktiv avatarların yaradılması potensial təhlükə yaradır [14].

2. Modelin qurulması - Bu mərhələdə müxtəlif süni intellekt modelləri inteqrasiya olunaraq tam funksional sintetik avatar formalaşdırılır. Vizual komponent üçün GAN və diffusion modelləri istifadə olunaraq üzün yüksək realizm səviyyəsində sintezi həyata keçirilir [4].

Səs komponentində isə "voice cloning" texnologiyaları tətbiq olunur və istifadəçinin səsi yüksək dəqiqliklə imitasiya edilir [19]. Böyük dil modellərinin inteqrasiyası nəticəsində isə sistem real-vaxt rejimində kontekstual cavablar verə bilir [17].

Bu mərhələnin əsas fərqləndirici xüsusiyyəti ondan ibarətdir ki, yaradılan sistem statik deyil, dinamikdir. Yəni avatar istifadəçi ilə interaktiv əlaqə qurur, suallara cavab verir və davranışını uyğunlaşdırır. Bu xüsusiyyət klassik Deepfake hücumlarından əsas fərq hesab olunur [6]. Deepfake əsaslı interaktiv avatarların praktik tətbiqini daha aydın göstərmək üçün HeyGen platforması vasitəsilə yaradılmış nümunə Şəkil 3-də təqdim olunur.

3. Hücumun icrası və biometrik sistemlərin bypass edilməsi - Son mərhələdə yaradılmış avatar autentifikasiya sisteminə real istifadəçi qismində təqdim olunur. Bu zaman hücum yalnız vizual uyğunluğa deyil, həm də davranış uyğunluğuna əsaslanır.

Müasir sistemlərdə tətbiq olunan canlılığın aşkarlanması (liveness detection) mexanizmləri (göz qırpma, baş hərəkəti və s.) interaktiv avatarlar tərəfindən real-vaxt rejimində imitasiya oluna bilir. Bu isə ənənəvi saxtalaşdırmaya qarşı müdafiə (anti-spoofing) metodlarının effektivliyini azaldır [8], [13].

Autentifikasiya sistemlərində istifadə olunan vektor təsviri (embedding) əsaslı müqayisə metodları sintetik məlumatlarla yüksək uyğunluq göstərə bilər. Bu yanaşmada istifadəçinin sifət və ya səs məlumatları çoxölçülü vektorlar şəklində qurulur və identifikasiya prosesi bu vektorlar arasındakı oxşarlığın ölçülməsi ilə həyata keçirilir. Bu məqsədlə geniş istifadə olunan kosinus oxşarlığı ölçüsü aşağıdakı kimi ifadə olunur:

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

burada A real istifadəçiyə, B isə təqdim olunan (məsələn, Deepfake) məlumata uyğun vektor təsviridir. Əgər bu oxşarlıq əvvəlcədən müəyyən edilmiş həddi (threshold) aşarsa, sistem autentifikasiyanı uğurlu hesab edir. Bu səbəbdən yüksək keyfiyyətli Deepfake sistemləri vektor təsviri məkanında real istifadəçiyə yaxın nəticələr yaradaraq autentifikasiya mexanizmlərini aldada bilər [20]

Şəkil 3. HeyGen platforması vasitəsilə yaradılmış interaktiv avatar nümunəsi.

Bu hücumların daha təhlükəli tərəfi onların sosial mühəndislik ilə birləşdirilməsidir. İnteraktiv avatar istifadəçi ilə dialoq quraraq əlavə məlumat əldə edə, istifadəçini manipulyasiya edə və hətta çoxfaktorlu autentifikasiya mexanizmlərindən dolayı yolla yan keçə (bypass) bilər [1].

Eyni zamanda, interaktiv avatarların real-vaxt rejimində cavab verə bilməsi onların klassik canlılığın aşkarlanması mexanizmlərini də keçməsinə imkan yaradır. Bu sistemlər adətən göz qırpma, baş hərəkəti və digər canlılıq göstəricilərinə əsaslansa da, interaktiv Deepfake sistemləri bu davranışları dinamik şəkildə imitasiya edə bilir [13].

V. MÜDAFİƏ STRATEGİYALARI VƏ MƏHDUDİYYƏTLƏR

Deepfake texnologiyalarına qarşı müdafiə mexanizmlərinin inkişafı bu sahədəki hücum texnologiyalarının inkişaf tempi ilə müqayisədə daha yavaş irəliləyir. Bu vəziyyət praktik mühitdə hücum edən tərəfin üstünlük qazanmasına və "hücum-müdafiə balansı"nın pozulmasına səbəb olur [21]. Başqa sözlə, generativ modellərin sürətli inkişafı nəticəsində daha realistik sintetik məlumatlar yaradıldığı halda, bu məlumatları aşkar edən sistemlər eyni sürətlə təkmilləşə bilmir.

Mövcud Deepfake aşkarlama yanaşmaları əsasən vizual və statistik xüsusiyyətlərin analizinə əsaslanır. Bu metodlar videolarda və şəkillərdə mövcud olan kiçik uyğunsuzluqları, məsələn piksel səviyyəsində fərqləri, kompressiya artefaktlarını və ya hərəkət ardıcıllığında qeyri-dəqiqlikləri müəyyən etməyə çalışır [22]. Bununla yanaşı, bəzi müasir yanaşmalar Konvolyusiya Neyron Şəbəkələri (Convolutional Neural Network, CNN), EfficientNet və ya transformer əsaslı modellərdən istifadə edərək daha yüksək səviyyəli xüsusiyyətləri analiz edir [23]. Lakin generativ modellər inkişaf etdikcə bu xüsusiyyətləri də imitasiya etməyi öyrənir və nəticədə klassik aşkarlama metodlarının effektivliyi tədricən azalır [2].

Bu isə daha adaptiv və öyrənə bilən müdafiə sistemlərinə ehtiyac yaradır.

Real-vaxt rejimində Deepfake aşkarlanması sahəsində də müəyyən məhdudiyyətlər mövcuddur. Məsələn, yüksək dəqiqlikli modellər adətən daha çox hesablama resursu tələb edir və bu da onların real-vaxt rejimində tətbiqini çətinləşdirir. Digər tərəfdən, aşağı gecikməli modellər isə bəzən dəqiqlik baxımından kifayət qədər effektiv olmur. Bu səbəbdən real-vaxt sistemlərində dəqiqlik və performans arasında balansın qorunması əsas problemlərdən biri hesab olunur [24].

Bu problemlərin aradan qaldırılması üçün son dövrlərdə multimodal təhlükəsizlik yanaşmaları daha geniş tətbiq olunmağa başlamışdır. Bu yanaşmada yalnız bir məlumat növünə əsaslanmaq əvəzinə, müxtəlif mənbələrdən əldə olunan məlumatlar birlikdə analiz edilir. Məsələn, istifadəçinin sifət görüntüsü və səsi ilə yanaşı, onun davranış xüsusiyyətləri, reaksiyaları, cavab vermə sürəti və ümumi ünsiyyət konteksti də nəzərə alınır [7]. Bu cür kompleks analiz hücum edən tərəfin sistemləri aldatmasını daha çətinləşdirir və ümumi təhlükəsizlik səviyyəsini artırır.

Bununla yanaşı, yalnız texniki həllər bu problemin tam həlli üçün kifayət etmir. Təşkilati səviyyədə görülən tədbirlər də mühüm rol oynayır. İşçilərin Deepfake və sosial mühəndislik hücumları barədə məlumatlandırılması, mütəmadi təhlükəsizlik təlimlərinin keçirilməsi və istifadəçilərin diqqət səviyyəsinin artırılması bu risklərin azaldılmasına kömək edir [14]. Eyni zamanda kritik əməliyyatlar üçün əlavə təsdiq mexanizmlərinin tətbiqi, məsələn çoxfaktorlu autentifikasiya və ya iki mərhələli təsdiq prosesləri, bu tip hücumların qarşısını almaqda effektiv vasitə kimi çıxış edir.

Bununla belə, mövcud müdafiə strategiyalarının texniki baxımdan daha da dərinləşdirilməsinə ehtiyac vardır. Xüsusilə real-vaxt rejimində Deepfake aşkarlanması üçün istifadə olunan konkret alqoritmlərin strukturu, təlim metodları və performans göstəricilərinin daha ətraflı təhlili gələcək tədqiqatlar üçün vacib istiqamət hesab olunur. Bu cür yanaşma təqdim olunan həllərin praktik tətbiq imkanlarını daha aydın göstərməyə imkan verə bilər.

Ümumilikdə, Deepfake texnologiyalarına qarşı effektiv müdafiə yalnız bir metodla deyil, müxtəlif texniki və təşkilati tədbirlərin inteqrasiyası ilə mümkündür. Bu sahədə davamlı inkişaf, adaptiv alqoritmlərin tətbiqi və yeni yanaşmaların hazırlanması gələcəkdə təhlükəsizlik səviyyəsinin artırılması üçün əsas rol oynayacaqdır.

NƏTİCƏ

Deepfake əsaslı interaktiv avatarlar müasir informasiya mühitində yeni və mürəkkəb təhlükə kateqoriyası formalaşdırır. Bu texnologiyalar yalnız texniki sistemlərin zəif tərəflərindən deyil, həm də insan faktorundan istifadə edir. Xüsusilə vizual və səs məlumatlarına olan etibar bu sistemlərin daha inandırıcı olmasına və hücumların effektivliyinin artmasına səbəb olur.

Aparılan təhlil göstərir ki, mövcud biometrik autentifikasiya sistemləri bu tip hücumlara qarşı tam etibarlı deyil. Üz və səs əsaslı identifikasiya sistemləri bəzi hallarda sintetik məlumatlarla aldadıla bilir. Bu səbəbdən real-vaxt analiz, davranış əsaslı autentifikasiya və multimodal təhlükəsizlik yanaşmaları xüsusi əhəmiyyət kəsb edir.

Bununla yanaşı, Deepfake texnologiyalarının hüquqi və etik təsirləri də nəzərə alınmalıdır. Kimliyin saxtalaşdırılması və dezinformasiya kimi risklər bu sahədə tənzimləmə və etik prinsiplərin gücləndirilməsini zəruri edir.

Nəticə etibarilə, Deepfake texnologiyalarının yaratdığı risklərin azaldılması üçün texniki, təşkilati və hüquqi tədbirlərin inteqrasiyası vacibdir.

ƏDƏBİYYAT

[1] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," ACM Computing Surveys, 2021.

[2] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey," Information Fusion, 2020.

[3] I. Goodfellow et al., "Generative adversarial nets," in Advances in Neural Information Processing Systems (NeurIPS), 2014.

[4] T. Karras, S. Laine, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[5] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in International Conference on Learning Representations (ICLR), 2014.

[6] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Neural voice puppetry," in European Conference on Computer Vision (ECCV), 2020.

[7] T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2019.

[8] S. Marcel, M. Nixon, and S. Z. Li, Handbook of Biometric Anti-Spoofing. Springer, 2021.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems (NeurIPS), 2020.

[10] R. Rombach, A. Blattmann, Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[11] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in Advances in Neural Information Processing Systems (NeurIPS), 2021.

[12] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[13] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI-created fake videos by detecting eye blinking," in Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), 2018.

[14] R. Chesney and D. Citron, "Deepfakes and the new disinformation war," Foreign Affairs, 2019.

[15] T. C. Wang, M. Y. Liu, J. Y. Zhu, et al., "Video-to-video synthesis," in Advances in Neural Information Processing Systems (NeurIPS), 2020.

[16] Y. Ren et al., "FastSpeech 2," in International Conference on Learning Representations (ICLR), 2020.

[17] T. B. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems (NeurIPS), 2020.

[18] OpenAI, "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[19] X. Tan, T. Qin, F. Soong, and T. Y. Liu, "Neural speech synthesis survey," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021.

[20] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[21] L. Verdoliva, "Media forensics and deepfake detection," IEEE Signal Processing Magazine, 2020.

[22] A. Rössler, D. Cozzolino, L. Verdoliva, et al., "FaceForensics++: Learning to detect manipulated facial images," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[23] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning (ICML), 2019.

[24] T. T. Nguyen et al., "Deep learning for deepfakes detection," IEEE Access, 2022.

Security Analysis of Deepfake-Based Interactive Avatars and Their Risks for Biometric Authentication Systems

Ziyafat Amirov[1], Majid Gurbanov[2], Abbas Huseynzade[3]
[1,2,3] Heydar Aliyev Academy of the State Security Service, Baku, Azerbaijan
[1]zamirov@inbox.ru, [2]majidgurbanov@gmail.com

Abstract- Recent advancements in generative artificial intelligence have significantly expanded the capabilities of deepfake technologies, enabling the development of AI-driven interactive systems. These systems integrate visual, audio, and behavioral features to produce highly realistic real-time imitations, posing serious threats to biometric authentication mechanisms. This paper provides a systematic analysis of the architecture of deepfake-based interactive systems, their attack models, and their impact on biometric identification algorithms. Furthermore, existing detection methods and multimodal defense approaches are evaluated. The findings highlight the necessity of adaptive and multi-layered security strategies.

Keywords- deepfake; deepfake; digital avatar; generative adversarial network; variational autoencoder.