TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETI HUZURIDAGI ILMIY DARAJALAR BERUVCHI DSc.13/05.05.2023.T.07.03 RAQAMLI ILMIY KENGASH

TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETI

NASIMOVA NIGORAXON MIZROBOVNA

STATISTIK MA’LUMOTLAR ASOSIDA SINTETIK O‘QUV TANLANMANI HOSIL QILISHNING MODEL VA ALGORITMLARI

05.01.11 – Raqamli texnologiyalar va sun’iy intellekt

TEXNIKA FANLARI BO‘YICHA FALSAFA DOKTORI (PHD) DISSERTATSIYASI AVTOREFARATI

Toshkent – 2025

UDK: 004.93

Texnika fanlari bo‘yicha falsafa doktori (PhD) dissertatsiyasi avtoreferati mundarijasi

Оглавление автореферата диссертации доктора философии (PhD) по техническим наукам

Contents of the Doctor of Philosophy (PhD) in Technical Sciences Dissertation Abstract

Nasimova Nigoraxon Mizrobovna
Statistik ma’lumotlar asosida sintetik o‘quv tanlanmani hosil qilishning model va algoritmlari 3

Насимова Нигорахон Мизробовна
Модели и алгоритмы для генерации синтетических обучающих выборок на основе статистических данных 18

Nasimova Nigoraxon Mizrobovna
Models and algorithms of generating the synthetic dataset based on statistical data 39

E’lon qilingan ishlar ro‘yxati
Список опубликованных работ
List of published works 40

TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETI HUZURIDAGI ILMIY DARAJALAR BERUVCHI DSc.13/05.05.2023.T.07.03 RAQAMLI ILMIY KENGASH

TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETI

NASIMOVA NIGORAXON MIZROBOVNA

STATISTIK MA’LUMOTLAR ASOSIDA SINTETIK O‘QUV TANLANMANI HOSIL QILISHNING MODEL VA ALGORITMLARI

05.01.11 – Raqamli texnologiyalar va sun’iy intellekt

TEXNIKA FANLARI BO‘YICHA FALSAFA DOKTORI (PHD) DISSERTATSIYASI AVTOREFARATI

Toshkent – 2025

Texnika fanlari bo‘yicha falsafa doktori (PhD) dissertatsiyasi mavzusi O‘zbekiston Respublikasi Oliy ta’lim, fan va innovatsiyalar vazirligi huzuridagi Oliy attestatsiya komissiyasida B2025.1.PhD/T5320 raqam bilan ro‘yxatga olingan.

Dissertatsiya Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universitetida bajarilgan.

Dissertatsiya avtoreferati uch tilda (o‘zbek, rus, ingliz (rezyume)) Ilmiy kengash veb-sahifasida (www.tuit.uz) va “ZiyoNet” axborot ta’lim portalida (www.ziyonet.uz) joylashtirilgan.

Ilmiy rahbar:	**Mo‘minov Bahodir Boltaevich** texnika fanlari doktori, professor
Rasmiy opponentlar:	**Muxamediyeva Dildora Kabilovna** texnika fanlari doktori, dotsent
	Atadjanov Ibragim Ravshanbekovich texnika fanlari doktori
Yetakchi tashkilot:	**“Toshkent irrigatsiya va qishloq xo‘jaligini mexanizatsiyalash muhandislari instituti” milliy tadqiqot universiteti**

Dissertatsiya himoyasi Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti huzuridagi DSc.13/05.05.2023.T.07.03 raqamli Ilmiy kengashning 2025-yil “25”-iyun soat 14⁰⁰ dagi majlisida bo‘lib o‘tadi. (Manzil: 100084, Toshkent shahri, Amir Temur shox ko‘chasi 108-uy. Tel.: (99871) 238-64-43, e-mail: ilmiy_kengash@tuit.uz).

Dissertatsiya bilan Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universitetining Axborot-resurs markazida tanishish mumkin (357 raqam bilan ro‘yxatga olingan). (Manzil: 100084, Toshkent shahri, Amir Temur shox ko‘chasi 108-uy. Tel.: (99871) 238-64-70).

Dissertatsiya avtoreferati 2025-yil “12” iyun da tarqatildi.
(2025-yil “12” iyun dagi 9 raqamli reestr bayonnomasi.)

M.M.Kamilov
Ilmiy darajalar beruvchi ilmiy kengash raisi, texnika fanlari doktori, professor, O‘zR FA akademigi

N.A.Egamberdiyev
Ilmiy darajalar beruvchi ilmiy kengash ilmiy kotibi, texnika fanlari bo‘yicha falsafa doktori

N.O. Raximov
Ilmiy darajalar beruvchi ilmiy kengash qoshidagi ilmiy seminar raisi, texnika fanlari doktori, dotsent

KIRISH (Falsafa doktori (PhD) dissertatsiyasi annotatsiyasi)

Dissertatsiya mavzusining dolzarbligi va zarurati. Jahonda sun’iy intellekt algoritmlarini tibbiyotda, jumladan surunkali kasalliklarga chalinish xavfini baholash, kasallikning yashirin belgilarini aniqlash, rivojlanish tendensiyasini baholashda qo’llash yetakchi o’rinlardan birini egallamoqda. Ammo chuqur o‘qitish algoritmlarini o’qitish uchun katta hajmdagi o‘quv tanlanma talab etiladi. Dunyo miqyosida zaruriy o‘quv tanlanmaning yetishmovchiligi yoki ulardan foydalanish uchun Sog‘liqni saqlash sug‘urtasi portativligi va javobgarligi to‘g‘risidagi qonuni va Umumiy ma’lumotlarni himoya qilish reglamenti kabi qonuniy cheklovlarning mavjudligi haqiqiy o‘quv tanlanmalarning o’rniga sintetik o‘quv tanlanmalardan foydalanishni amaliyotga joriy etishni taqozo etadi[1]. Shu jihatdan ishonchli va sifatli sintetik tibbiy o‘quv tanlanmasini hosil qilish muhim ahamiyatga ega hisoblanadi.

Jahonda sintetik tibbiy o‘quv tanlanmalarni hosil qilish, ularning sifatini va ishonchliligini oshiruvchi yangi ilmiy-amaliy yechimlarini ishlab chiqishga yo’naltirilgan ilmiy-tadqiqot ishlari olib borilmoqda. Bu borada, sintetik ma’lumotlarni hosil qilish imkonini beruvchi Synthea, MDClone Adam, Gretel, Synthetic Data Vault kabi dastur va kutubxonalarni ishlab chiqishga alohida e’tibor berilmoqda. Gartner Amerika texnologik tadqiqotlar markazi sintetik ma’lumotlarni ishlab chiqarish bozori hajmi 2024 va 2029-yillarda 61,1% ga ya’ni 4,39 milliard dollarga o‘sadi deb bashorat qilgan[2]. Shuning uchun bugungi kunda jahon miqyosida statistik ma’lumotlar asosida sifatli va ishonchli sintetik o‘quv tanlanmalarni hosil qilish dolzarb ilmiy tadqiqot masalalaridan biri bo’lib qolmoqda.

Respublikamizda tibbiyot yo‘nalishida tibbiy ma’lumotlar bazasini to‘plash va samaradorlikni oshirish uchun ularni sintetik hosil qilish imkonini beruvchi dasturiy vositalarni ishlab chiqish yuzasidan keng qamrovli chora-tadbirlar amalga oshirilib, muayyan natijalarga erishilmoqda. 2030-yilgacha O‘zbekiston Respublikasida sun’iy intelektni yanada rivojlantirish bo’yicha Harakatlar strategiyasida, jumladan, “… katta hajmdagi ma’lumotlar to‘plamini tahlil qilish va bilimlarni to‘plash bo‘yicha, shu jumladan katta hajmdagi ma’lumotlarni yig‘ish, saqlash va intellektual tahlil qilishning yangi usullari va algoritmlarini, katta hajmdagi ma’lumotlarni tarqatish uchun yangi usullar va dasturlar, shu bilan birga murakkab muhandislik yechimlarini bashoratli modellashtirish uchun yangi usullar va dasturiy ta’minotlar bo‘yicha ilmiy-tadqiqot ishlarini olib borish” bo’yicha muhim vazifalar belgilab berilgan. Ushbu vazifalarini amalga oshirishda, jumladan, statistik ma’lumotlar asosida keng tarqalgan kasalliklarga chalingan bemorlar haqida yetarlicha ma’lumotlarni o’zida mujassamlashtirgan sifatli va ishonchli sintetik jadvalli o’quv tanlanmalarni hosil qilish imkonini beruvchi sun’iy intellekt model va algoritmlarini yaratish muhim ahamiyat kasb etmoqda.

O‘zbekiston Respublikasi Prezidentining 2020-yil 5-oktyabrdagi PF-6079-sonli ““Raqamli O‘zbekiston - 2030” strategiyasini tasdiqlash va uni samarali amalga oshirish chora-tadbirlari to‘g‘risida”gi farmonida[3], 2023–yil 11– sentabrdagi

[1] Fonseca, J., Bacao, F. Tabular and latent space synthetic data generation: a literature review. J Big Data 10, 115 (2023).

[2] https://www.technavio.com/report/synthetic-data-generation-market-analysis

[3] O‘zbekiston Respublikasi Prezidentining 05.10.2020 yildagi PF-6079-sonli, ““Raqamli O‘zbekiston — 2030” strategiyasini tasdiqlash va uni samarali amalga oshirish chora-tadbirlari to‘g‘risida”gi farmoni.

PF–158–sonli “O‘zbekiston–2030 strategiyasi to‘g‘risida”gi farmonida[4], O‘zbekiston Respublikasi Prezidentining 2024-yil 14-oktyabrdagi “Sun’iy intellekt texnologiyalarini 2030-yilga qadar rivojlantirish strategiyasini tasdiqlash to‘g‘risida”gi PQ-358-sonli qarorlari[5] hamda mazkur faoliyatga tegishli boshqa me’yoriy-huquqiy hujjatlarda belgilangan vazifalarni amalga oshirishga ushbu dissertatsiya ishi muayyan darajada xizmat qiladi.

Tadqiqotning respublika fan va texnologiyalari rivojlanishining ustuvor yo‘nalishlariga mosligi. Mazkur tadqiqot respublika fan va texnologiyalari rivojlanishining IV. “Axborotlashtirish va axborot-kommunikatsiya texnologiyalarini rivojlantirish” ustuvor yo‘nalishi doirasida bajarilgan.

Muammoning o‘rganilganlik darajasi. So‘nggi yillarda jahonda sintetik jadvalli ma’lumotlarni sun’iy intellekt model va algoritmlari asosida hosil qilish va amaliyotda joriy etishga doir ilmiy izlanishlar olib borilmoqda. Jumladan, L.Xu, M.Skularidu, A.Kuesta-Infant, Y.Choy, S. Bisval, B.Malin, J.Duk, A.Mottini, N.Park, M.Mohammadi, K.Gord kabi tadqiqotchilarning ishlarida Generativ teskari tarmoqlari (Generative Adversarial Networks, GANs) yordamida sintetik jadvalli ma’lumotlarni hosil qilishning turli usullari taklif etilgan. Variatsion avtoankoder (Variational autoencoder, VAE) tarmoqlari yordamida sintetik jadvalli ma’lumotlarni hosil qilish usullari N.Park, M.Mohammadi, K.Gord tadqiqotlarida ko‘rib chiqilgan, mashinali o‘qitish usullari, xususan Bayes tarmog‘i yordamida sintetik jadvalli ma’lumotlarni hosil qilish usullari G.Zang, K.Kormod, D.Prokopiuk kabi olimlar tominidan ishlab chiqilgan.

Respublikamizda ma’lumotlarga intellektual ishlov berish sohasining tanib olish model va algoritmlari yo‘nalishida M.Kamilov, xususiyatlarni aniqlash va tanlash algoritmlari yo‘nalishida Sh.Fozilov, ekspert tizimlari yo‘nalishida M.Raxmatullayev, tasniflash va tashxislash masalalari bo‘yicha A.Nishanov, sintetik ma’lumotlarni hosil qilishning model va algoritmlari yo‘nalishida B.Mo‘minov, F.Maxmudxo‘jayev, Sh.Madraximov kabi olimlar ilmiy tadqiqot olib borib, ushbu yo‘nalishning rivoji uchun o‘z hissalarini qo‘shib kelishmoqda.

Shu bilan birga, statistik ma’lumotlarga asoslangan sintetik tibbiy jadvalli ma’lumotlar bazasini hosil qilish yo‘nalishida tadqiqotlarni davom ettirish dolzarb bo‘lib qolmoqda.

Dissertatsiya tadqiqotining dissertatsiya bajarilgan oliy ta’lim muassasasining ilmiy-tadqiqot ishlari rejalari bilan bog‘liqligi. Dissertatsiya tadqiqoti Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universitetining ilmiy-tadqiqot ishlari rejasining И3-2020082913 “Surunkali kasalliklarni masofadan monitoring qilish uchun smartfon dasturiy ilovasini ishlab chiqish” (2022-2023) mavzusidagi loyiha doirasida bajarilgan.

Tadqiqotning maqsadi surunkali kasalliklarni sun’iy intellekt bilan bashoratlashda ishlatiladigan sintetik o‘quv tanlanmalarni statistik ma’lumotlar asosida hosil qilish model va algoritmlarini ishlab chiqishdan iborat.

[4] O‘zbekiston Respublikasi Prezidentining 11.09.2023 yildagi PF-158-sonli, “O‘zbekiston–2030 strategiyasi to‘g‘risida”gi farmoni

[5] O‘zbekiston Respublikasi Prezidentining 14.10.2024-yildagi PQ-358-sonli “Sun’iy intellekt texnologiyalarini 2030-yilga qadar rivojlantirish strategiyasini tasdiqlash to‘g‘risida” gi qarori.

Tadqiqotning vazifalari:

sintetik jadvalli o‘quv tanlanmalarni hosil qilishning sun’iy intellekt model va algoritmlarini tadqiq etish;

GAN tarmog‘ini o‘qitishda ishlatiladigan yo‘qotish funksiyalarini tahlil qilish asosida jadvallar yaqinligini baholovchi yangi yo‘qotish funksiyasining matematik modelini ishlab chiqish;

qisman ikkilik inversiya usuli asosida “0” va “1” dan iborat bo‘lgan jadvallarning qiymatlarini ustunlar bo‘yicha aralashtirish algoritmini ishlab chiqish;

mustahkamlab o‘qitish asosida sintetik jadvalli ma’lumotlarni hosil qilish algoritmini ishlab chiqish;

GAN tarmog‘ining kirish va chiqish qatlamlarini modifikatsiyalash va yangi yo‘qotish funksiyasini ishlab chiqish asosida sintetik o‘quv tanlanmalarni hosil qilishning model va algoritmi ishlab chiqish.

taklif etilayotgan model va algoritmlar asosida ishlab chiqilgan sintetik jadvallarni o‘xshashlilik va foydalilik mezonlari bo‘yicha baholash;

Tadqiqotning obyekti sifatida surunkali kasalliklarning statistik ma’lumotlari, GAN tarmog‘i, yo‘qotish funksiyasi, mustahkamlab o‘qitishning model va algoritmini ishlab chiqish jarayonlari qaralgan.

Tadqiqotning predmeti sifatida sintetik o‘quv tanlanmalarni hosil qilish model, usul va algoritmlari olindi.

Tadqiqotning usullari sifatida chuqur o‘qitishning algoritmlari, matematik va statistik tahlil usullari, neyron tarmoqlarini modellashtirish usullari olingan.

Tadqiqotning ilmiy yangiligi quyidagilardan iborat:

sintetik jadvalli ma’lumotlarning statistik xususiyatlarini inobatga olgan holda ularning haqiqiy ma’lumotlarga yaqinligini baholovchi va GAN tarmog‘ini o‘qitish uchun yangi yo‘qotish funksiyasining matematik modeli ishlab chiqilgan;

qisman ikkilik inversiya usuli asosida “0” va “1” dan iborat bo‘lgan jadvallarning ustunlaridagi qiymatlarni aralashtirish algoritmi ishlab chiqilgan;

mustahkamlab o‘qitish asosida o‘zgaruvchilarni aralashtirib sintetik jadvalli o‘quv tanlanmani hosil qilish algoritmi ishlab chiqilgan;

GAN tarmog‘ining kirish va chiqish qatlamlarini modifikatsiyalash hamda yangi ishlab chiqilgan yo‘qotish funksiyasini qo‘llash asosida sintetik o‘quv tanlanmalarni hosil qilishning model va algoritmi ishlab chiqilgan.

Tadqiqotning amaliy natijalari quyidagilardan iborat:

tadqiqot natijasida ishlab chiqilgan model va algoritmlar kesimida ma’lum fan soha mutaxassisi, dasturchi bo‘lmagan foydalanuvchilar uchun mo‘ljallangan dastur ishlab chiqilgan;

“Tibbiy sintetik ma’lumotlarni hosil qilish algoritmi” hamda “Tibbiy statistik ma’lumotlar asosida sintetik jadvalli ma’lumotlarni hosil qilishning GAN algoritmi” nomli algoritmlar ishlab chiqilgan;

“Surunkali kasalliklarni masofadan monitoring qilishning shifokor ilovasi” hamda “Surunkali kasalliklarni masofadan monitoring qilishning bemor ilovasi” nomli dasturiy ilovalar ishlab chiqilgan.

Tadqiqot natijalarining ishonchliligi. Tadqiqot natijalarining ishonchliligi tavsiya etilgan baholash metodlari (Student t-testi, Vassershteyn masofasi, χ^2 (Chi-kvadrat) testi, TSTR (Train on Synthetic, Test on Real) va TRTR (Train on Real, Test on Real) testlari) natijalari bilan tasdiqlangan. Qo‘yilgan muammoning matematik jihatdan to‘g‘ri ifodalanishi, xususiyatlar orasidagi bog‘liqliklarni aniqlashda ma’lumotlarni intellektual tahlil qilish usullari to‘g‘ri qo‘llanilgan.

Tadqiqot natijalarining ilmiy va amaliy ahamiyati. Tadqiqot natijalarining ilmiy ahamiyati taklif etilgan neyron tarmog‘i arxitekturasi, sintetik o‘quv tanlanmalarini hosil qilish algoritmi ko‘plab boshqa surunkali kasalliklarni tashxislash va bashoratlash bilan bog‘liq masalalarini hal etishda foydalanish bilan izohlanadi.

Tadqiqot natijalarining amaliy ahamiyati hosil qilingan o‘quv tanlanma asosida surunkali kasalliklarni monitoring qiluvchi mobil ilova orqali bemorlarni o‘z-o‘zini monitoring qilish va mas’ul shifokorlar masofadan monitoring qilishi imkonini hosil qilish orqali surunkali kasalliklarning og‘ir asoratlarini oldini olish, shuningdek 2-tur qandli diabet xavfini oldindan aniqlash imkonini hosil qilish bilan izohlanadi.

Tadqiqot natijalarining joriy qilinishi. Yaratilgan tadqiqot usullari va algoritmlari bo’yicha olingan natijalar asosida:

“Tibbiy statistik ma’lumotlar asosida sintetik jadvalli ma’lumotlarni hosil qilishning GAN algoritmi” asosida hosil qilingan sintetik o’quv tanlanmadan foydalangan holda yaratilgan 8 ta simptom asosida 2-tur diabetga tashxis qo‘yish imkoniyatini beruvchi dasturiy vositasi Olot tuman tibbiyot birlashmasiga qarashli ko‘p tarmoqli markaziy poliklinikasida muvaffaqiyatli joriy etildi. Ushbu dastur qandli diabet kasalliklarini tashxislashda aniqlikni oshirishga xizmat qilib, uning amaliyotga tatbiq etilishi (O‘zbekiston Respublikasi Raqamli texnologiyalar vazirligining 2025-yil 25-fevraldagi 33-8/1255-sonli ma’lumotnomasi) bo‘yicha samarali deb baholandi. Joriy qilingan dastur soha mutaxassislarining ish unumdorligini 5 % ga oshirgan;

“Tibbiy sintetik ma’lumotlarni hosil qilish algoritmi” asosida hosil qilingan sintetik o‘quv tanlanmalar yordamida yaratilgan yurak xurujini aniqlash imkoniyatini beruvchi dasturiy vosita Marg‘ilon tumanlararo Perinatal markazida muvaffaqiyatli joriy etildi. Ushbu dastur qandli diabet kasalliklarini tashxislashda aniqlikni oshirishga xizmat qilib, uning amaliyotga tatbiq etilishi (O‘zbekiston Respublikasi Raqamli texnologiyalar vazirligining 2025-yil 25-fevraldagi 33-8/1255-sonli ma’lumotnomasi) bo‘yicha samarali deb baholandi. Joriy qilingan dastur soha mutaxassislarining ish unumdorligini 7% ga oshirgan;

“Tibbiy sintetik ma’lumotlarni hosil qilish algoritmi” hamda “Tibbiy statistik ma’lumotlar asosida sintetik jadvalli ma’lumotlarni hosil qilishning GAN algoritmi” dasturlari tibbiyot xodimlarining kasbiy malakasini rivojlantirish markazida joriy etilgan (O‘zbekiston Respublikasi Raqamli texnologiyalar vazirligining 2025-yil 25-fevraldagi 33-8/1255-sonli ma’lumotnomasi). Ushbu ishlab chiqilgan algoritm va dasturlar yordamida olingan natijalar axborot texnologiyalari va sun’iy intellekt sohasida ta’lim berish jarayonida talabalarga o‘quv tanlanmalar bilan ishlash, ulardan algoritmlarni o‘qitishda foydalanish, mavjud xatoliklarni aniqlash, muayyan

masalani yechishda haqiqiy o‘quv tanlanma topilmaganda sintetik o‘quv tanlanmadan foydalanish foydalari va kamchiliklarini tushunishlarida muhim rol o‘ynab, ta’lim samaradorligini oshirish imkoniyatini bergan.

Tadqiqot natijalarining aprobatsiyasi. Dissertatsiyaning ilmiy va amaliy natijalari 9 ta xalqaro va 5 respublika ilmiy-amaliy anjumanlarida muhokamadan o‘tkazilgan.

Tadqiqot natijalarining e’lon qilinganligi. Tadqiqot mavzusi bo‘yicha asosiy natijalari 28 ta ilmiy ishlarda e’lon qilingan, ulardan 7 tasi O‘zbekiston Respublikasi Oliy attestatsiya komissiyasi tomonidan doktorlik dissertatsiyalarining asosiy ilmiy natijalarini e’lon qilish uchun tavsiya qilingan jurnallarda, jumladan 5 tasi xorijiy jurnallarda va 2 tasi respublika jurnallarida nashr qilingan, hamda 5 ta EHM uchun yaratilgan dasturiy vositalarni qayd qilish guvohnomalari olingan.

Dissertatsiyaning tuzilishi va hajmi. Dissertatsiya to‘rtta bob, xulosa, foydalanilgan adabiyotlar ro‘yxati va ilovadan iborat. Dissertatsiyaning hajmi 118 betni tashkil etadi.

DISSERTATSIYANING ASOSIY MAZMUNI

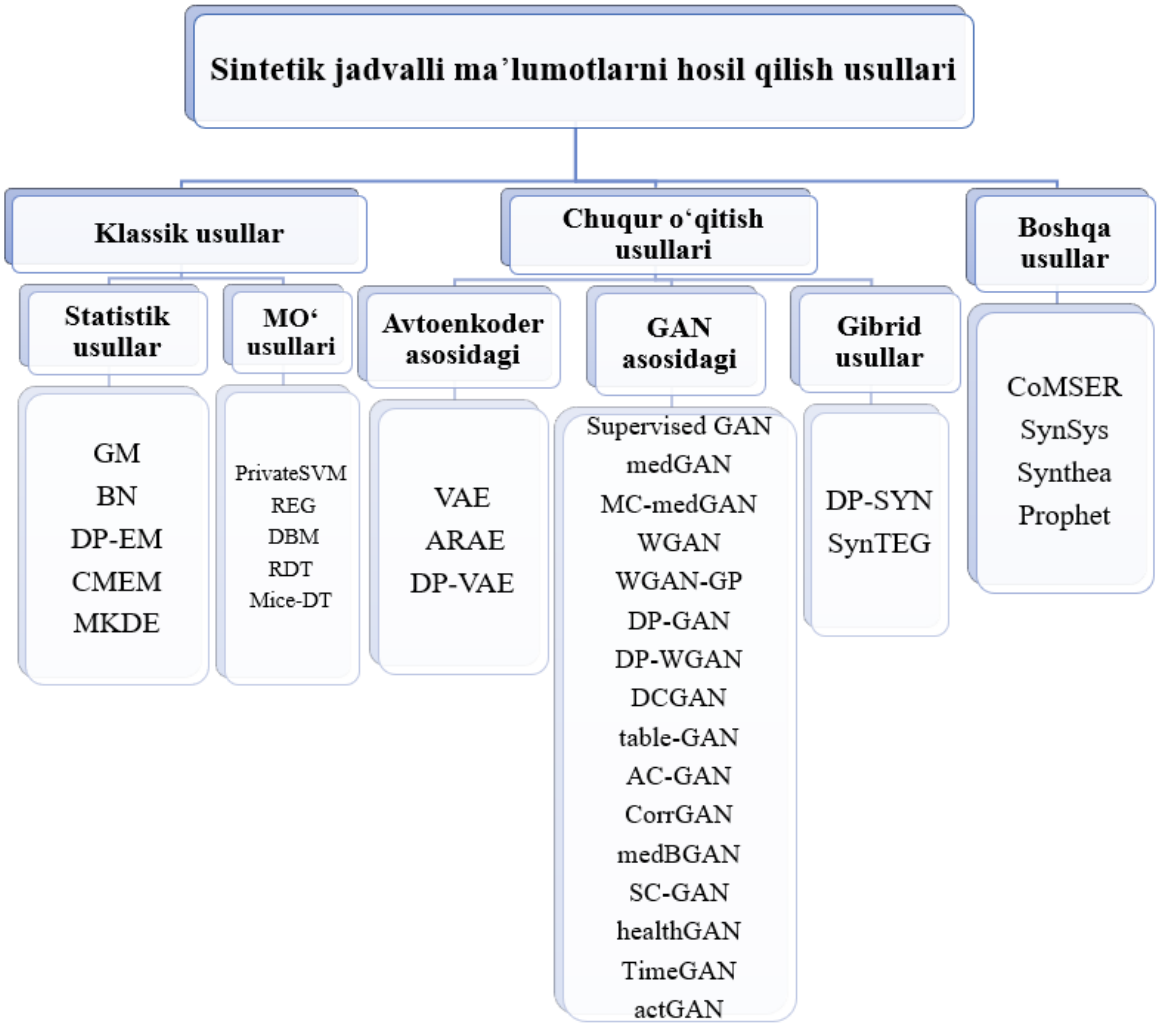
Kirish qismida O‘zbekiston Respublikasining fan va texnologiyalar taraqqiyotining ustuvor yo‘nalishlariga muvofiq dissertatsiya mavzusining dolzarbligi va zaruriyati asoslangan, tadqiqotning maqsadi va vazifalari shakllantirilgan hamda tadqiqot obyekti va predmeti ko‘rsatilgan, tadqiqotning ilmiy yangiliklari hamda amaliy natijalari keltirilgan, olingan natijalaraning ishonchligi asoslangan, nazariy va amaliy ahamiyati ko‘rsatilgan, tadqiqot natijalarini amaliyotga joriy qilinishi, nashr etilgan ishlar va dissertatsiya ishining tuzilishi bo‘yicha ma’lumotlar keltirilgan.

Dissertatsiyaning **“Sintetik ma’lumotlarning ahamiyati va ularni hosil qilish algoritmlari tahlili” deb nomlangan birinchi bobida** sintetik ma’lumotlarning sun’iy intellektni tibbiyotda qo‘llash jarayonidagi o‘rni va ahamiyati, sintetik o‘quv tanlanmalarining qo‘llanilish sohalari, jadvalli sintetik o‘quv tanlanmalarini hosil qilishdagi muammolar va dolzarb masalalar o‘rganilgan. Ushbu sohada dunyo tadqiqotchilari tomonidan olib borilayotgan ilmiy va amaliy tadqiqot ishlari atroflicha o‘rganilib, sintetik jadvalli o‘quv tanlanmalarni hosil qilish usuli va algoritmlari qiyosiy tahlil qilingan. Tadqiqotlar natijalaridan hozirda ushbu sohada hal etilishi dolzarb bo‘lgan masalalar keltirilgan.Sintetik ma’lumot tushunchasiga Qirollik jamiyati va Alan Tyuring instituti tomonidan “Ma’lumotlar ilmi fani (Data science)ga oid vazifalarni hal qilish maqsadida maxsus tuzilgan, matematik model yoki algoritm yordamida hosil qilingan ma’lumotlardir” deb ta’rif beriladi. Sintetik ma’lumotlar qator imkoniyatlarni taqdim etadi.

Sintetik jadvalli ma’lumotlarlarni hosil qilish murakkab jarayon bo‘lib, bu undagi bir nechta muammolar bilan bog‘liqdir.

- Ma’lumotlarning aralash turli bo‘lishi.
- Ma’lumotlarning ishonchliligi.
- Domenga xoslik.

– Yechiladigan masalaga xoslik.
– Rejimning qulashi.
– Statistik o‘xshashlikning talab etilishi.
– Ma’lumotlar maxfiyligini ta’minlash murakkabligi.
– Kam sonli ma’lumotlar bilan ishlash muammosi.



1-rasm. Sintetik ma’lumotlarni hosil qilish usullari tasnifi

Ushbu muammolarga qaramasdan, sintetik jadvalli ma’lumotlarga bo‘lgan ehtiyoj sababli so‘nggi yillarda sintetik jadvalli ma’lumotlarni hosil qilish bo‘yicha qator tadqiqotlar olib borildi. Natijada, sintetik tibbiy ma’lumotlar bazasini hosil qilishning bir qancha model va algoritmlari ishlab chiqildi (1-rasm). Ammo ular orasida statistik ma’lumotlardan foydalanib sintetik o‘quv tanlanmalarni hosil qilish usullari juda kam va mavjudlari ham SI algoritmlari hisoblanmaydi, balki statistik usullar hisoblanadi. Ulardan foydalanishda hisoblash murakkabligi, foydalanuvchidan chuqur matematik va statistik bilimlarni talab etilishi kabi muammolarga duch kelish mumkin. Shu sababli statistik ma’lumotlar asosida SI algoritmlari, xususan chuqur o‘qitish algoritmlari yordamida jadvalli sintetik o‘quv tanlanmalarni hosil qilish dolzarb vazifa hisoblanadi.

Dissertatsiyaning **“GAN tarmog‘ini o‘qitishda yo‘qotish funksiyalari” deb nomlangan ikkinchi bobida** GAN tarmog‘ini o‘qitishda yo‘qotish funksiyasining roli, keng tarqalgan yo‘qotish funksiyalari turlari va ularning xususiyatlari,

shuningdek, turli yo‘qotish funksiyalarining GAN tarmog‘i hosil qilgan ma’lumotlar sifatiga ta’siri, ularning afzallik va kamchiliklari o‘rganilgan. Bundan tashqari, jadvalli ma’lumotlarning o‘xshashligini baholash uchun ishlatiladigan funksiyalar, ularning xususiyatlari qiyosiy tahlil qilingan. Ushbu funksiyalardan neyron tarmoqlarini o‘qitishda yo‘qotish funksiyasi sifatida foydalanish imkoniyatlari qiyosiy tahlillar asosida ko‘rib chiqilgan.

Har bir yo‘qotish funksiyasi o‘zining kuchli va zaif tomonlariga ega, shuning uchun GAN tarmog‘ini optimallashtirishda turli yo‘qotish funksiyalarining kombinatsiyasi yaxshiroq natijalarga olib kelar ekan.

Jadvallarning o‘xshashligini baholovchi yo‘qotish funksiyalari sintetik ma’lumotlarni hosil qilishda muhim rol o‘ynaydi, chunki ular haqiqiy va sintetik jadvallar orasidagi farqlarni aniqlashga yordam beradi. Asosiy maqsad sintetik jadvallarni haqiqiy ma’lumotlarga maksimal darajada o‘xshash qilish. Lekin bugungi kunda ishlatiladigan yo‘qotish funksiyalari qator kamchiliklarga ega, jumladan, jadvallarni sifatli baholay olmaslik, hisoblashning murakkabligi, o‘qitish uchun ishlatish noqulay.

Dissertatsiya ishining **uchinchi bobi “Sintetik ma’lumotlarni hosil qilish algoritmlari”** deb nomlangan bo‘lib, unda yangi yo‘qotish funksiyasining matematik modeli, GAN tarmog‘i asosida sintetik ma’lumotlarni hosil qilishning yangi algoritmi, faqat 0 va 1 qiymatlardan iborat bo‘lgan jadvalli ma’lumotlarni hosil qilishga mo‘ljallangan mustahkamlab o‘qitish algoritmi ishlab chiqilgan.

Jadvallarning o‘xshashligini baholash uchun yangi yo‘qotish funksiyasi ishlab chiqilgan bo‘lib, uni matematik modeli quyidagicha:

1-qadam. Har bir satr uchun yig‘indini hisoblash: Haqiqiy jadval *R* matritsa, sintetik jadval *S* matritsa sifatida berilgan bo‘lsin. Bunda, x_{nm}- *R* ning *n*-qator va m-ustunidagi element, y_{nm} esa S matritsaning i-qatori va j-ustunidagi elementdir. Bizning holatda matritsalar 766 qator va 9 ustundan iborat. n-qatordagi *R* matritsa elementlarining va i-qatordagi S matritsa elementlarining yig‘indisi hisoblanadi:

$$r_n = \sum_m^9 x_{nm} \qquad (n=1,2,\ldots,9) \qquad (1)$$

$$s_i = \sum_j^9 y_{ij} \qquad (i=1,2,\ldots,9) \qquad (2)$$

Yuqoridagi yig‘indini har bir *n* va *i* qator uchun bajariladi.

2-qadam. Har bir ustun uchun yig‘indini hisoblash: *R* matritsaning *m*-ustunidagi elementlarining va *S* matritsaning *j*-ustunidagi elementlarining yig‘indisi hisoblanadi:

$$R_m = \sum_n^{766} x_{nm} \qquad (m=1,2,\ldots,766) \qquad (3)$$

$$S_j = \sum_i^{766} y_{ij} \qquad (j=1,2,\ldots,766) \qquad (4)$$

*3-qadam.*Vertikal ayirmalarning o‘rtacha qiymatini aniqlash: Har bir ma’lumotlarning ustunlari yig‘indisi elementlar bo‘yicha ayirilib, ayirmalarning o‘rtacha qiymati hisoblanadi: C={$\left| R_m - S_j \right|$}.

$$\mu = \frac{1}{9}\sum_{k=1}^{9} C_k \qquad (5)$$

4-qadam.Har bir qator uchun ayirmalarning minimal qiymati hisoblanadi:

$$\delta_n = \min_i(|r_n - s_i|) \quad (6)$$

Natijada δ_n minimal qiymatlardan iborat M to‘plam hosil bo‘ladi: $M = \{\delta_1, \delta_2, \dots, \delta_n\}$, bunda, *n= 1, 2, …766*.

5-qadam. Qayta takrorlanadigan minimal qiymatning foizini topish: Bunda, *N* teng qiymatga ega bo‘lgan 766 ta minimumga ega hamda takrorlangan minimal qiymat foizi aniqlanadi: P = N/766

6-qadam. Minimal qiymatlarning o‘rtacha qiymatini aniqlash:

$$\nu = \frac{1}{766}\sum_{n=1}^{766}\delta_n \quad (7)$$

7-qadam. Gorizontal yo‘qotishlarni hisoblash:

$$D = \frac{\nu}{P} \quad (8)$$

8-qadam. Bitta namuna uchun umumiy yo‘qotish quyidagicha aniqlanadi:

$$L_{bn} = D + \mu \quad (9)$$

Tarmoqni o‘qitish davomida har bir batch uchun L_{1bn}, L_{2bn} ,…, L_{kbn} yo‘qotish qiymatlariga ega bo‘lamiz, Bunda *k* - to‘plamdagi misollar soni.

9-qadam. Umumiy yo‘qotish funksiyasi barcha batchlardan olingan yo‘qotish qiymatlarining kvadrat ildizi orqali hisoblaniladi:

$$L_{TD} = \sqrt{\frac{1}{k}\sum_{l=1}^{k} L_l} \quad (10)$$

Mustahkamlab o‘qitish algoritmi 2-rasmda tasvirlangan. Ushbu algoritm quyidagi bosqichlarni o‘z ichiga oladi.

1-qadam. Algoritmga statistik ma’lumotlarni kiritish.

2-qadam. Ushbu statistik ma’lumotlar va ularga mos formulalar asosida birlamchi ma’lumotlar bazasi ishlab chiqish.

3-qadam. Fisher-Yats algoritmi yordamida ma’lumotlarni ustunlar bo‘yicha aralashtirish.

4-qadam. Taklif etilgan yangi aralashtirish algoritmi yordamida ma’lumotlarni ustunlar bo‘yicha aralashtirish.

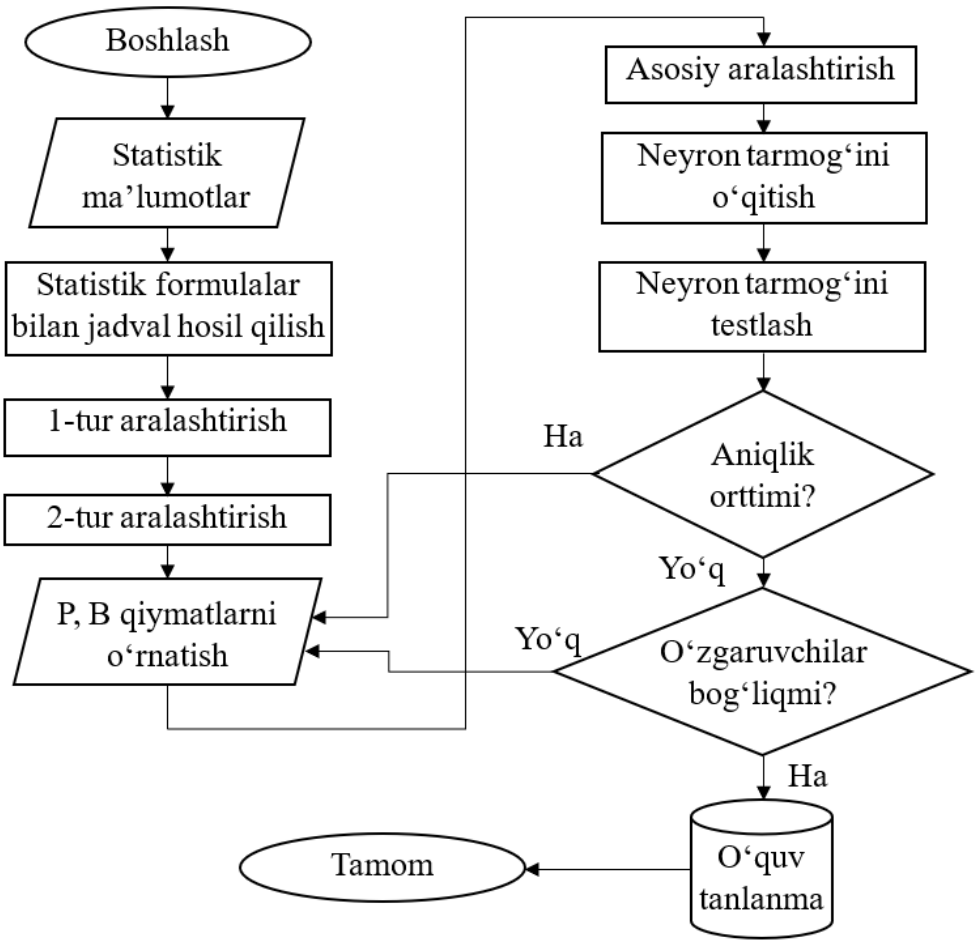
5-qadam. P va B koeffitsiyentlarning boshlang’ich qiymatlarini va ularni o‘zgarishini aniqlashda ishlatiladigan β_1, β_2, ε koeffitsiyentlarning qiymatlarini kiritish.

6-qadam. Fisher-Yats algoritmi va *P, B* koeffitsiyentlardan foydalangn holda ma’lumotlarni belgilangan oralig’ini ustunlar bo‘yicha aralashtirish.

7-qadam. Aralashtirilgan o‘quv tanlanmani neyron tarmog‘iga uzatish va unda 5 davr davomida o‘qitish.

8-qadam. Sintetik o‘quv tanlanmada o‘qitilgan neyron tarmog‘ini haqiqiy o‘quv tanlanmada o‘qitib sinab ko‘rish.

9-qadam. Aniqlik yetarli bo‘lsa, o‘quv tanlanmani saqlab qo‘yish.

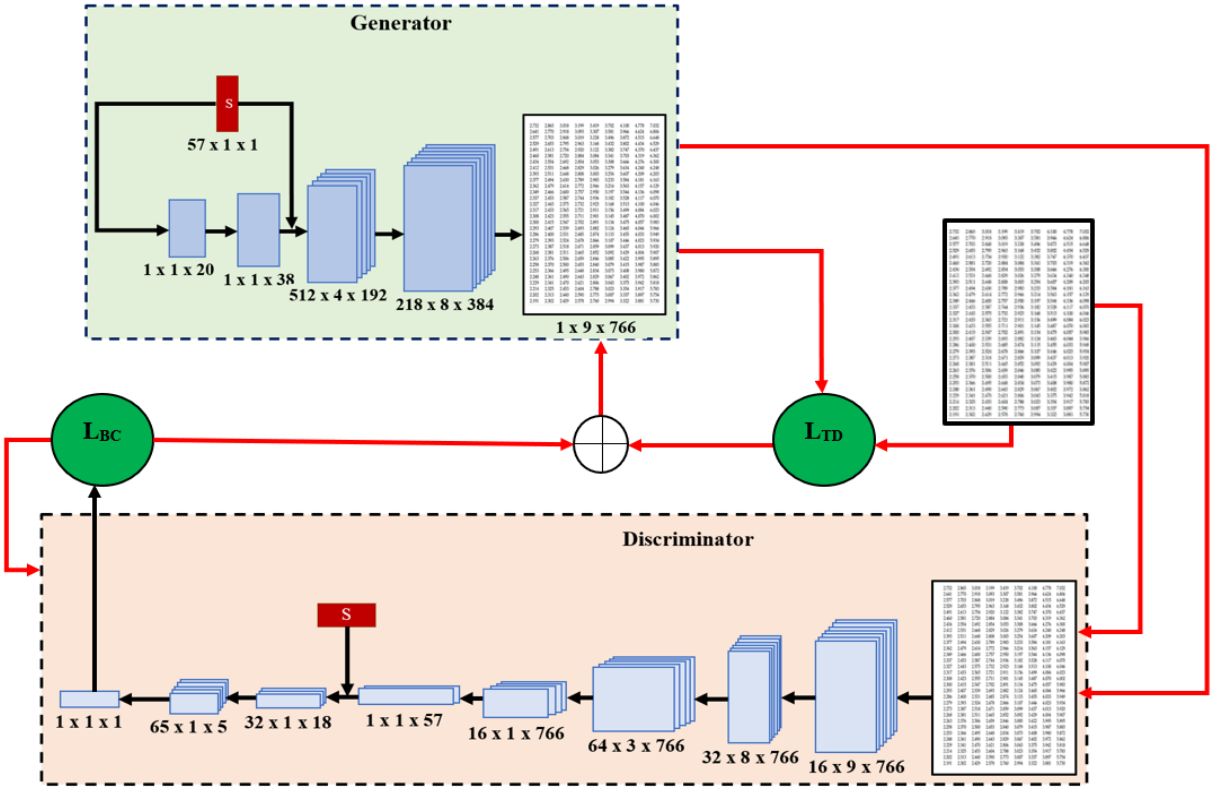


2-rasm. Berilgan statistik ma’lumotlar asosida sintetik ma’lumotlar bazasi hosil qilish algoritm

An’anaviy GAN tarmog‘i ikki qismdan iborat bo‘ladi: generator va diskriminator. Ushbu ikkita neyron tarmog‘i o‘zaro raqobatlashib birgalikda o‘qitiladi. Generator tarmog‘i - tasodifiy shovqinni kiruvchi ma’lumot sifatida qabul qiladi va uni sintetik ma’lumotlar namunalariga aylantiradi. Generatorning muvaffaqiyatli ishlashi uning diskriminatorni aldashi mumkin bo‘lgan yuqori sifatli va bir-birini takrorlamaydigan namunalarni hosil qilish qobiliyati bilan baholanadi. Generator yuqori generatsiya qobiliyatiga yo‘qotish funksiyasini minimallashtirish orqali erishadi. Diskriminator tarmog‘ining vazifasi esa generator ishlab chiqqan namunalarni haqiqiylik ehtimolligini baholashdan iboratdir. Tarmoqni o‘qitish davomida diskriminator ma’lumotlar to‘plamidagi haqiqiy ma’lumotlar va generator tomonidan hosil qilingan sun’iy namunalarni farqlashni o‘rganadi.

An’anaviy GAN modelini qo‘yilgan maqsadga moslashtirish uchun unga quyidagi uchta o‘zgartirishlar kiritildi: birinchidan, generator kirishini bitta emas, ikkita qilindi va kirishga tasodifiy shovqin emas, statistik ma’lumotlar kiritildi, ikkinchidan, generator tarmog‘ini o‘qitish uchun yangi yo‘qotish funksiyasidan foydalanildi, uchinchidan, diskriminator chiqishidan bir necha qatlam oldin statistik ma’lumotlar kiritildi. Ushbu modifikatsiyalangan GAN tarmog‘ining generator qismi ConvTranspose, BatchNormalization, Dropout, ReLu, Flatten va Output qatlamlaridan, Diskriminator qismi esa Convolution, BatchNormalization, Dropout, ReLu, Flatten va Output qatlamlaridan tashkil topgan.

GAN tarmog‘i yordamida sintetik ma’lumotlarni hosil qilish uchun an’anaviy GAN tarmog‘ining modifikatsiyalangan modeli ishlab chiqildi (3-rasm).



3-rasm. Taklif etilgan GAN tarmog‘i arxitekturasi

Generator va diskriminatorning o‘zaro raqobati natijasida, umumiy GAN tarmog‘i haqiqiy malumotlarga juda oxshash sintetik ma’lumotlarni hosil qilishni boshlaydi.

Tegishli yo‘qotish funksiyasi mos ravishda diskriminator uchun quyidagi:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[lgD\left(x^{(i)}\right) + \lg(1 - D(G(z^{(i)}))) \right] \qquad (11)$$

Generator uchun esa quyidagi:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} lg(1 - D(G(z^{(i)}))) + L_{TS} \qquad (12)$$

ifoda orqali aniqlanadi. Bunda, L_{TS} – qo‘shimcha yo‘qotish funksiyasi.

GAN tarmog‘i statistik ma’lumotlarga asosan jadval hosil qilishni o‘rganadi, shuning uchun bu tarmoqni o‘qitishda zarur bo‘lgan ma’lumotlar bazasi ikkitadan iborat bo‘ladi: jadvalli ma’lumotlar bazasi va ushbu jadvallarning tegishli statistik xususiyatlari aks etgan statistik ma’lumotlar bazasi. Bu jarayonning bajarilish ketma-ketligi 4-rasmda tasvirlangan.

1-qadam. Statistik ma’lumotlar bazasini hosil qilish.

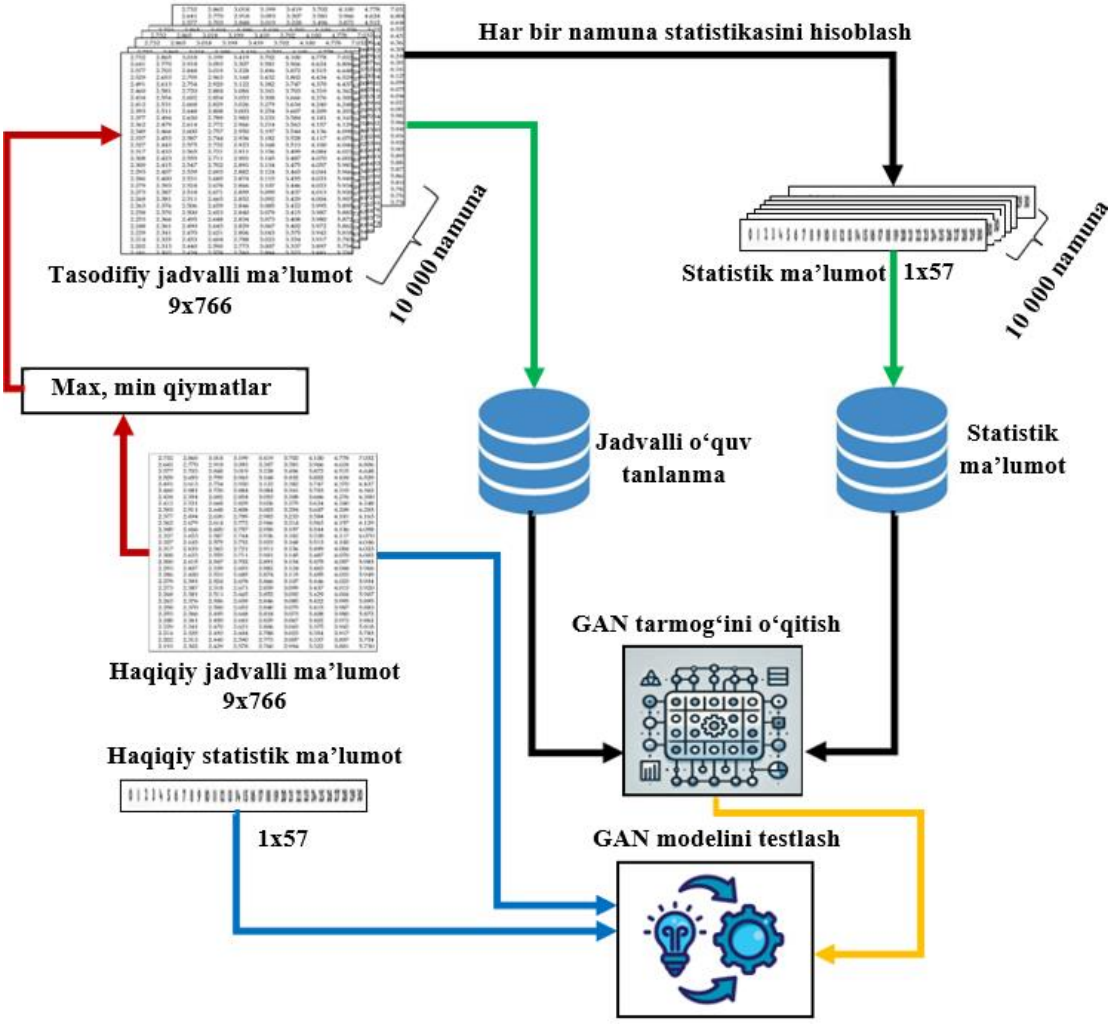
2-qadam. Statistik ma’lumotlar bazasining har bir ma’lumoti uchun birlamchi jadvalli ma’lumotlarni hosil qilish. Ushbu jadvalli ma’lumotlarni birlashtirgan holda jadvalli ma’lumotlar bazasini hosil qilish.

3-qadam. GAN tarmog‘ini hosil qilingan ikkita (jadvalli va statistik) ma’lumotlar bazasi bilan o‘qitish. O‘qitish davomida hosil qilingan sintetik o‘quv

tanlamaning o‘zgaruvchilari orasidagi bog‘liqlik haqiqiy o‘quv tanlanmaning o‘zgaruvchilari orasidagi bog‘liqlik bilan solishtirib boriladi.

4-qadam. O‘qitilgan GAN tarmog’i haqiqiy statistik ma’lumotlar asosida hosil qilgan sintetik o‘quv tanlanmalar foydalilik, o‘xshashlik mezonlari bo‘yicha baholanadi.

5-qadam. Baholash natijasi qoniqarli deb hisoblangan GAN tarmog‘i arxitekturasining giperparametrlari saqlab qo‘yiladi.



4-rasm. GAN tarmog‘i yordamida sintetik o‘quv tanlanmani hosil qilish algoritmi

Dissertatsiyaning **“Sun’iy intellekt algoritmlari yordamida hosil qilingan sintetik o‘quv tanlanmalarni baholash”** deb nomlangan to‘rtinchi bobida sintetik jadvalli ma’lumotlarning sifatini baholash mezonlari va ularning tasnifi, ta’riflari keltirilgan. Yangi model va algoritmlar yordamida hosil qilingan sintetik jadvalli o‘quv tanlanmalarning haqiqiy jadvalli o‘quv tanlanmalarga o‘xshashligi va sun’iy intellekt algoritmlarini o‘qitish uchun foydaliligi tanlangan mezonlar asosida baholangan. Sintetik ma’lumotlar hosil qilingandan so‘ng, uning sifatini baholash eng muhim vazifa hisoblanadi. Ushbu tadqiqot ishida sintetik jadvalli o‘quv tanlanma (SJO‘T) ni kompleks baholash uchun quyidagi usullar ajratib olindi:

1. SJO‘T ning o‘xshashligini baholash uchun tanlangan usullar: Student T-test, Chi-square ($\chi 2$) test, Vasseshteyn masofasi.

2. SJO‘T ning foydaliligini baholash uchun tanlangan usullar: TRTR va TSTR testlardan olingan aniqlik, F1-qiymat, o‘ziga xoslik, sezgirlik qiymatlari va ular orasidagi farqlar.

O‘xshashlikni baholash uchun Vasseshteyn masofasidan foydalanildi. Bunda o‘quv tanlanmaning har bir atributi uchun Vassershteyn masofasi alohida-alohida hisoblandi. 1-jadvalda olingan natijalar keltirilgan. Jadvaldan ko‘rinib turibdiki, taklif qilingan usul yordamida hosil qilingan o‘quv tanlanmalardagi barcha atributlar uchun Vassershteyn masofasi kritik qiymat (0,3) dan kam bo‘lgan, bu esa ushbu attributlar o‘xshashlik shartini bajarishini ko‘rsatadi.

1-jadval.

Vassershteyn masofasi qiymati.

Atributlar nomi	*Taklif etilgan usul*	*GM*	*CTGAN*
Homiladorlik soni	0.077049	0.050503	0.090343
Glukoza miqdori	0.139552	0.029401	0.211049
Qon bosimi	0.119209	0.077486	0.135793
Teri qalinligi	0.166620	0.082580	0.063013
Insulin	0.047145	0.381628	0.161786
MVN	0.155688	0.026149	0.111652
Diabet Pedigre Funksiyasi	0.176867	0.051625	0.057591
Yoshi	0.086909	0.022935	0.031314

2-jadval.

TSTR testi natijalari

	Baholash mezonlari			
Usul nomi	*Aniqlik (%)*	*Sezgirlik (%)*	*O‘ziga xoslik (%)*	*F1-qiymat (%)*
Haqiqiy o‘quv tanlanma	0.7532	0.7554	0.7532	0.7542
SDV	0.6494	**0.7648**	0.6494	0.6494
CTGAN	0.3312	0.388	0.3312	0.2812
Taklif etilgan usul	**0.7078**	0.6966	**0.7078**	**0.6926**

Foydalilikni baholash. Foydalilikni baholash uchun TSTR testidan foydalanildi. Bu testda SI modeli sifatida Random Forest (RF) modelidan foydalanildi. Har bir model uchun test natijalarining aniqlik, o‘ziga xoslik, sezgirlik va F1-qiymatlari hisoblandi. 2-jadvalda TSTR test natijalari haqidagi ma’lumotlar keltirilgan.

XULOSA

“Statistik ma’lumotlar asosida sintetik o‘quv tanlanmalarni hosil qilishning model va algoritmlari” mavzusidagi dissertatsiya bo‘yicha olib borilgan tadqiqotlar natijasida quyidagi xulosalar olindi:

1. Sintetik jadvalli o‘quv tanlanmalarni hosil qilishning dolzarb masalalari va muammolar, sintetik jadvalli tibbiy o‘quv tanlanmalarni hosil qilishning usullari qiyosiy tahlil qilindi. Sintetik o‘quv tanlanmalarni hosil qilish usullarini solishtirma tahlil qilish natijasida turli surunkali kasalliklarga chalinish xavfini oldindan bashoratlash uchun ishlatiladigan jadvalli sintetik o‘quv tanlanmalarga ehtiyoj yuqori ekanligi asoslandi.

2. Sintetik o‘quv tanlanmalar GAN tarmog‘i yordamida hosil qilish jarayonida ishlatiladigan yo‘qotish funksiyalari tasniflanib, ularning xususiyatlari solishtirma tahlil qilindi. Bunda jadvalli sintetik o‘quv tanlanmalarni hosil qilishning usullari juda murakkab ekanligi va aynan jadvallarning o‘xshashligini baholash uchun moslashmaganligi sababli jadvalli ma’lumotlarning yaqinligini baholash uchun maxsus yo‘qotish funksiyasini ishlab chiqish dolzarb ekanligi aniqlandi.

3. Yuqoridagi talabdan kelib chiqib, to‘la sintetik tibbiy ma’lumotlarni faqat statistik ma’lumotlardan va neyron tarmog‘idan foydalangan holda mustahkamlab o‘qitish usulida hosil qilish algoritmi ishlab chiqildi.

4. Sintetik jadvalli o‘quv tanlanmalarni GAN tarmog‘i yordamida hosil qilishda faqat bitta yo‘qotish funksiyasidan foydalanish yaxshi samara bermasligi aniqlandi. Bunda ushbu asosiy yo‘qotish funksiyasiga qo‘shimcha yo‘qotish funksiyasini qo‘shish tarmoq samaradorligini oshirishi aniqlandi.

5. Jadvalli sintetik tibbiy o‘quv tanlanmalarni hosil qilishda GAN tarmog‘idan foydalanishda ishlatish mumkin bo‘lgan yangi yo‘qotish funksiyasi ishlab chiqildi. Ushbu y o‘qotish funksiyasi GAN va VAE tarmoqlarini o‘qitishda hosil qilinuvchi jadvalning sifatini yaxshilash imkonini beradi.

6. Sintetik jadvalli ma’lumotlarni GAN tarmog‘i yordamida hosil qilish algoritmi ishlab chiqildi. Ushbu algoritm statistik ma’lumotlar asosida kerakli jadvalli o‘quv tanlanmalarni hosil qilish imkonini beradi.

7. Statistik o‘quv tanlanmalarga asosan 2-tur diabetga chalinish xavfini 5 yil oldindan bashoratlash sun’iy intellekt algoritmlarini o‘qitish uchun foydalanish mumkin bo‘lgan o‘quv tanlanma ishlab chiqildi. Ushbu o‘quv tanlanma yordamida o‘qitilgan neyron tarmog’i 2-tur diabetga chalinish xavfini 5 yil oldindan bashoratlash samaradorligini 5% ga oshirgan.

8. Modifikatsiya qilingan GAN tarmog‘i yordamida hosil qilingan sintetik tanlanmalar ishonchliligi va haqiqiy ma’lumotlarga o‘xshashligi, sun’iy intetllekt modellarini o‘qitishda foydaliligi nuqtai nazaridan yaxshi sifatli ma’lumotlar ekanligi turli baholash mezonlari bo‘yicha aniqlandi.

НАУЧНЫЙ СОВЕТ DSc.13/05.05.2023.T.07.03 ПО ПРИСУЖДЕНИЮ УЧЕНЫХ СТЕПЕНЕЙ ПРИ ТАШКЕНТСКОМ УНИВЕРСИТЕТЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

ТАШКЕНТСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

НАСИМОВА НИГОРАХОН МИЗРОБОВНА

МОДЕЛИ И АЛГОРИТМЫ ДЛЯ ГЕНЕРАЦИИ СИНТЕТИЧЕСКИХ ОБУЧАЮЩИХ ВЫБОРОК НА ОСНОВЕ СТАТИСТИЧЕСКИХ ДАННЫХ

05.01.11 – Цифровые технологии и искусственный интеллект

АВТОРЕФЕРАТ ДИССЕРТАЦИИ
ДОКТОРА ФИЛОСОФИИ (PhD) ПО ТЕХНИЧЕСКИМ НАУКАМ

Ташкент – 2025

Тема диссертации доктора философии (PhD) по техническим наукам зарегистрирована в Высшей аттестационной комиссии при Министерство высшего образования, науки и инноваций Республики Узбекистан за B2025.1.PhD/T5320.

Диссертация выполнена в Ташкентском университете информационных технологий.

Автореферат диссертации на трех языках (узбекский, английский, русский (резюме)) размещен на веб странице (www.tuit.uz) и на Информационно-образовательном портале «Ziyonet» (www.ziyonet.uz).

Научный руководитель:	**Муминов Баходир Болтаевич** доктор технических наук, профессор
Официальные оппоненты:	**Мухамедиева Дилдора Кабиловна** доктор технических наук, доцент
	Атаджанов Ибрагим Равшанбекович доктор технических наук
Ведущая организация:	**Национальный исследовательский университет «Ташкентский институт инженеров ирригации и механизации сельского хозяйства»**

Защита диссертации состоится «25» июня 2025 г. в 14⁰⁰ часов на заседании научного совета DSc.13/05.05.2023.T.07.03 при Ташкентском университете информационных технологий. (Адрес: 100084, г. Ташкент, ул. Амира Темура, 108. Тел.: (99871) 238-64-43; e-mail: ilmiy_kengash@tuit.uz).

С диссертацией можно ознакомиться в Информационно-ресурсном центре Ташкентского университета информационных технологий (регистрационный номер №357). (Адрес: 100084, г. Ташкент, ул. Амира Темура, 108. Тел.: (99871) 238-64-70).

Автореферат диссертации разослан «12» июня 2025 года.
(протокол рассылки №9 от «12» июня 2025 г.).

М.М.Камилов
Председатель Научного совета по присуждению учёных степеней, доктор технических наук, академик АН РУз

Н.А.Эгамбердиев
Ученый секретарь Научного совета по присуждению учёных степеней, доктор философии по техническим наукам

Н.О.Рахимов
Председатель научного семинара при Научном совете по присуждению учёных степеней, доктор технических наук, доцент

ВВЕДЕНИЕ (аннотация диссертации доктора философии (PhD))

Актуальность и востребованность темы диссертации. На сегодняшний день проводится множество научных исследований по оценке риска хронических заболеваний с помощью алгоритмов искусственного интеллекта, в том числе алгоритмов глубокого обучения, выявлению скрытых признаков заболевания, оценке тенденции развития. Однако для обучения алгоритмам углубленного обучения требуется большой объем обучающей выборки. Одной из основных проблем остается отсутствие необходимой учебной выборки или наличие законодательных ограничений на их использование, таких как Закон о портативности и ответственности медицинского страхования и Общий регламент защиты данных. В последние годы исследователи предложили использовать синтетическую учебную выборку вместо реальной учебной выборки в качестве решения этой проблемы[1]. Поскольку качественная синтетическая учебная выборка сохраняет статистические характеристики и внутренние связи в реальной учебной выборке, они используются при обучении алгоритмам углубленного обучения. Однако создание качественных и надежных синтетических данных - очень сложная задача. Поэтому создание надежной и качественной синтетической медицинской образовательной выборки остается одной из актуальных задач.

В мире ведутся интенсивные исследовательские работы по созданию синтетических медицинских образовательных выборок, повышению их качества и надежности. В частности, в настоящее время все больше разрабатывается программ и библиотек, таких как Synthea, MDClone Adam, Gretel, Synthetic Data Vault, которые позволяют генерировать синтетические данные. Американский центр технологических исследований Gartner прогнозирует, что объем рынка производства синтетических данных вырастет на 61,1% в 2024 и 2029 годах, то есть на 4,39 миллиарда долларов[2]. Однако практически все они требуют использования существующей реальной учебной выборки и считаются неудовлетворительными для использования в медицине с точки зрения надежности и качества. Поэтому на сегодняшний день в мировом масштабе формирование качественных и достоверных синтетических учебных выборок на основе статистических данных остается одной из актуальных научно-исследовательских задач.

В нашей республике особое внимание уделяется разработке программных средств, позволяющих собирать медицинские базы данных и синтезировать их для повышения эффективности. Данное диссертационное исследование в определенной степени слу жит реализации задач, предусмотренных Указом Республики Узбекистан от 5 октября 2020 года № PF -6079 «Об утверждении Стратегии «Цифровой Узбекистан - 2030» и мерах по ее эффективной

[1]Fonseca, J., Bacao, F. Tabular and latent space synthetic data generation: a literature review. J Big Data 10, 115 (2023).

[2] https://www.technavio.com/report/synthetic-data-generation-market-analysis

реализации»[3], В Указе Президента Республики Узбекистан от 11 сентября 2023 года № PF -158 «О Стратегии «Узбекистан-2030» »[4] определены задачи, а именно «…развитию и стимулированию научно-исследовательских работ в области цифровых технологий, совершенствованию организационных механизмов их реализации…», а также по «…проведению научных исследований в области анализа и накопления знаний на основе больших объемов данных, в том числе по новым методам и алгоритмам сбора, хранения и интеллектуального анализа больших данных, новым способам и программам их распространения, а также по созданию новых методов и программного обеспечения для прогностического моделирования сложных инженерных решений»…», постановлением Президента Республики Узбекистан от 14 октября 2024 года №PQ-358 «Об утверждении Стратегии развития технологий искусственного интеллекта до 2030 года»[5], приоритетных задач «…внедрения технологий искусственного интеллекта: в сфере здравоохранения – определение методов диагностики, лечения заболеваний, анализ медицинских снимков и управление данными о больном...».

Соответствие исследования приоритетным направлениям развития науки и технологий республики. Диссертация выполнена в соответствии с приоритетным направлением развития науки и технологий республики IV. «Информатизациия и развитие информационно-коммуникационных технологий».

Степень изученности проблемы. В последние годы в мире проводятся научные исследования, посвященные генерации и внедрению в практику синтетических табличных данных на основе моделей и алгоритмов искусственного интеллекта. В частности, в работах таких исследователей, как L.Xu, M.Skularidu, A.Kuesta-Infant, Y.Choy, S. Bisval, B.Malin, J.Duk, A.Mottini, N.Park, M.Mohammadi, K.Gord, предложены различные методы генерации синтетических табличных данных с использованием генеративных состязательных сетей (Generative Adversarial Networks, GANs). Методы генерации синтетических табличных данных с помощью вариационных автоэнкодеров (Variational Autoencoder, VAE) рассмотрены в исследованиях N.Park, M.Mohammadi, K.Gord, а методы генерации синтетических табличных данных на основе методов машинного обучения, в частности с использованием байесовских сетей, разработаны такими учеными, как G.Zang, K.Kormod, D.Prokopiuk.

В Республике в сфере моделей и алгоритмов распознавания в области интеллектуальной обработки данных научные исследования ведет М. Камилов, в области алгоритмов выявления и отбора признаков – Ш. Фозилов, по экспертным системам – М. Рахматуллаев, по вопросам классификации и диагностики – А. Нишанов, а в сфере моделей и алгоритмов генерации синтетических данных – такие ученые, как Б. Муминов, Ф. Махмудходжаев и

[3] Указ Президента Республики Узбекистан от 05.10.2020 г. No УП-6079 «Об утверждении Стратегии «Цифровой Узбекистан-2030» и мерах по ее эффективной реализации».

[4] Указ Президента Республики Узбекистан от 11.09.2023 г. No УП-158 «О Стратегии «Узбекистан-2030»»

[5] Постановление Президента Узбекистан от 14 октября 2024 года №ПП-358 «Об утверждении Стратегии развития технологий искусственного интеллекта до 2030 года».ф

Ш. Мадрахимов, которые вносят свой вклад в развитие данного направления. Вместе с тем, продолжение исследований в области генерации базы синтетических медицинских табличных данных на основе статистических данных остается актуальной задачей.

Связь темы диссертационного исследования с планами научно-исследовательских работ высшего образовательного учреждения, где выполнена диссертация. Диссертационное исследование выполнено в соответствии с планом научно-исследовательских работ Ташкентского университета информационных технологий имени Мухаммада ал-Хоразмий в рамках проекта №ИЗ-2020082913 на тему: «Разработка программного приложения смартфона для дистанционного мониторинга хронических заболеваний» (2022-2023).

Целью исследования является разработка моделей и алгоритмов генерации синтетических обучающих выборок на основе статистических данных для прогнозирования хронических заболеваний с использованием искусственного интеллекта.

Задачи исследования:

исследование моделей и алгоритмов искусственного интеллекта для генерации синтетических табличных обучающих выборок;

разработка математической модели новой функции потерь, оценивающей близость таблиц, на основе анализа функций потерь, применяемых при обучении GAN-сетей;

разработка алгоритма перемешивания значений таблиц, состоящих из «0» и «1», по столбцам на основе метода частичной бинарной инверсии;

разработка алгоритма генерации синтетических табличных данных на основе обучения с подкреплением;

разработка алгоритма генерации синтетической табличной обучающей выборки на основе модифицированной GAN-сети;

оценка сгенерированных синтетических таблиц по критериям схожести и полезности на основе предлагаемых моделей и алгоритмов.

Объектом исследования являются статистические данные о хронических заболеваниях, GAN-сети, функции потерь, а также процессы разработки моделей и алгоритмов обучения с подкреплением.

Предметом исследования – модели, методы и алгоритмы генерации синтетических обучающих выборок.

Методы исследования. В ходе исследования использованы алгоритмы глубокого обучения, методы математического и статистического анализа, методы моделирования нейронных сетей.

Научная новизна исследования заключается в следующем:

разработана математическая модель функции потерь для обучения GAN-сетей, оценивающая близость синтетических табличных данных к реальным с учетом их статистических характеристик;

разработана алгоритм перемешивания значений в столбцах таблиц, состоящих из «0» и «1», на основе метода частичной бинарной инверсии;

разработан алгоритм генерации синтетических табличных обучающих выборок на основе обучения с подкреплением путем смешивания переменных;

разработаны модель и алгоритм формирования синтетических обучающих выборок на основе модификации входных и выходных слоев сети GAN и применения новый разработанной функции потерь.

Практические результаты исследования заключаются в следующем:

на основе моделей и алгоритмов, разработанных в результате исследования, разработана программа, предназначенная для пользователей, не являющихся программистами, специалистами в определенной области науки;

разработаны алгоритмы «Алгоритм генерирования медицинских синтетических данных» и «Алгоритм GAN генерирования синтетических табличных данных на основе медицинских статистических данных»;

разработаны программные приложения «Докторское приложение дистанционного мониторинга хронических заболеваний» и «Пациентное приложение дистанционного мониторинга хронических заболеваний» .

Достоверность результатов исследования подтверждена результатами применения предложенных методов оценки (t-критерий Стьюдента, расстояние Вассерштейна, χ^2 (хи-квадрат) тест, тесты TSTR (обучение на синтетических данных, тестирование на реальных) и TRTR (обучение и тестирование на реальных данных)), также корректной математической формализацией поставленной задачи и правильного применения методов интеллектуального анализа данных для выявления взаимосвязей между признаками.

Научная и практическая значимость результатов исследования. Научная значимость результатов исследования обусловлена тем, что предложенные архитектура нейронной сети и алгоритм генерации синтетических обучающих выборок могут быть применены для решения задач, связанных с диагностикой и прогнозированием широкого спектра хронических заболеваний.

Практическая значимость результатов исследования заключается в том, что на основе сгенерированной синтетической обучающей выборки возможна реализация мобильного приложения для мониторинга хронических заболеваний, которое позволяет пациентам осуществлять самостоятельный контроль состояния здоровья, а врачам – дистанционно наблюдать за пациентами. Это, в свою очередь, способствует предупреждению тяжелых осложнений хронических заболеваний и раннему выявлению риска развития сахарного диабета II типа.

Внедрение результатов исследования. На основе разработанных в рамках исследования методов и алгоритмов полученные результаты внедрены в практику по следующим направлениям:

компьютерная программа, позволяющий ставить диагноз сахарного диабета II типа по 8 симптомам, разработанная с использованием синтетических обучающих выборок, созданной на основе «Алгоритма GAN для формирования синтетических табличных данных на основе медицинской статистики» успешно внедрена в многопрофильной центральной

поликлинике, входящей в структуру медицинского объединения Алатского района. Данная программа способствует повышению точности диагностики заболеваний, связанных с сахарным диабетом, и ее внедрение в практику оценено как эффективное (справка Министерства цифровых технологий от 25 февраля 2025 года №33-8/1255). Внедренная система увеличила производительность труда специалистов отрасли на 5%;

компьютерная программа, позволяющий выявлять сердечный приступ по 12 заданным симптомам с использованием синтетических обучающих выборок, созданных на основе «Алгоритма генерации медицинских синтетических данных», успешно внедрена в Маргиланском межрайонном перинатальном центре. Данная программа способствует повышению точности диагностики заболеваний, связанных с сахарным диабетом, и его внедрение в практику признано эффективным (справка Министерства цифровых технологий от 25 февраля 2025 года №33-8/1255). Внедренная система повысила производительность труда специалистов на 7%;

программы «Алгоритм генерации синтетических медицинских данных» и «GAN-алгоритм генерации синтетических табличных медицинских данных на основе статистической информации» внедрены в Центре повышения квалификации медицинских работников (справка Министерства цифровых технологий от 25 февраля 2025 года №33-8/1255). Данные разработанные алгоритмы и программные решения сыграли важную роль в образовательном процессе в области информационных технологий и искусственного интеллекта. Они позволяют студентам работать с обучающими выборками, использовать их для обучения алгоритмов, выявлять ошибки и понимать, как синтетические выборки могут служить альтернативой реальным в случаях их отсутствия, а также осознавать их преимущества и ограничения, что способствует повышению эффективности образовательного процесса.

Апробация результатов исследования. Результаты настоящего исследования апробированы на 9 международных и 5 Республиканских научно-практических конференциях.

Публикация результатов исследования. По теме диссертации опубликовано 28 научных работ, в том числе 7 статей в изданиях, рекомендованных Высшей аттестационной комиссией Республики Узбекистан для публикации основных научных результатов диссертаций (из них 7 статьи – в республиканских журналах и 5 статей – в зарубежных журналах). Также получены 5 авторских свидетельств на программные средства для ЭВМ.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, заключения, списка использованной литературы и приложения. Объем диссертации составляет 118 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во **введении** обоснована актуальность и востребованность темы исследования, определена степень изученности проблемы, сформулированы

цель и задачи, объект и предмет исследования, раскрыто соответствие работы приоритетным направлениям развития науки и технологий в Республике Узбекистан, изложены научная новизна и практические результаты, определены теоретическая и практическая значимость полученных результатов, приведены сведения о внедрении результатов исследования в практику, апробации и опубликованных работах, структуре диссертации.

В первой главе диссертации «**Значение синтетических данных и анализ алгоритмов их генерации**» рассматриваются роль и значение синтетических данных в процессе применения искусственного интеллекта в медицине, области использования синтетических обучающих выборок, а также существующие проблемы и актуальные задачи при генерации табличных синтетических обучающих выборок. Подробно проанализированы научные и прикладные исследования, проводимые учеными по всему миру в данной области, выполнен сравнительный анализ методов и алгоритмов генерации синтетических табличных обучающих выборок. На основе анализа результатов этих исследований обозначены наиболее актуальные задачи, стоящие перед специалистами в данной сфере на сегодняшний день. Согласно определению, предложенному Королевским обществом и Институтом Алана Тьюринга, синтетические данные – это «информация, искусственно сгенерированная для решения задач в сфере науки о данных (Data science), сгенерированная с использованием математической модели или алгоритма». Синтетические данные предоставляют целый ряд возможностей.

Генерация синтетических табличных данных представляет собой сложный процесс, связанный с рядом проблем, в числе которых:

смешанный тип данных (наличие как числовых, так и категориальных признаков);

надежность данных (соответствие синтетических данных реальности);

доменно-специфическая зависимость (учет особенностей предметной области);

зависимость от конкретной задачи (специфика решаемой проблемы);

коллапс режима (mode collapse – проблема генеративных моделей, особенно GAN);

требование статистического сходства с реальными выборками;

сложность обеспечения конфиденциальности данных (особенно в медицинской сфере);

проблема работы с ограниченным объемом данных (малые обучающие выборки).

Несмотря на перечисленные проблемы, в последние годы, в связи с растущей потребностью в синтетических табличных данных, проведен ряд исследований по их генерации. В результате разработаны различные модели и алгоритмы генерации баз синтетических медицинских данных (см. рис. 1). Однако среди них крайне мало методов, основанных на использовании статистических данных для генерации синтетических обучающих выборок, и существующие методы не относятся к алгоритмам искусственного

интеллекта, а представляют собой статистические подходы. При их использовании могут возникнуть такие проблемы, как вычислительная сложность и необходимость глубоких знаний в области математики и статистики со стороны пользователя. В связи с этим задача генерации табличных синтетических обучающих выборок на основе статистических данных с использованием алгоритмов искусственного интеллекта, в частности алгоритмов глубокого обучения, является актуальной.

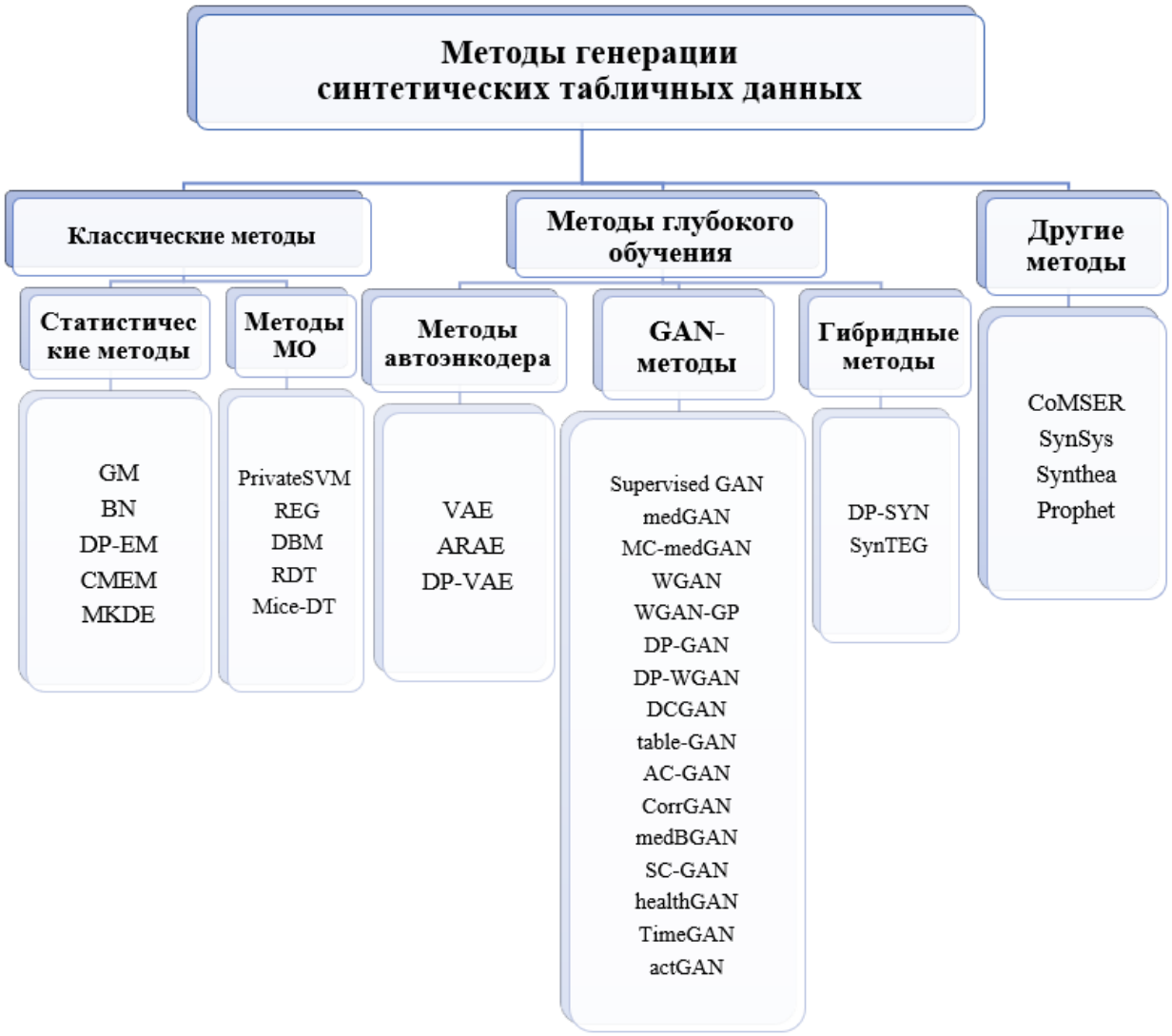


Рис. 1. Классификация методов генерации синтетических данных

Во второй главе диссертации «**Функции потерь при обучении GAN-сетей**» рассматривается роль функции потерь в обучении генеративных состязательных сетей (GAN), дается обзор распространенных типов функций потерь и их характеристик, а также анализируется влияние различных функций потерь на качество данных, генерируемых GAN-сетями. Также рассматриваются их преимущества и недостатки. Особое внимание уделено функциям оценки схожести табличных данных, их свойствам и возможностям использования в качестве функций потерь при обучении нейронных сетей. Возможности таких применений обоснованы на основе сравнительного анализа.

Каждая функция потерь имеет свои сильные и слабые стороны, поэтому комбинирование нескольких функций потерь при обучении GAN-сетей может привести к более оптимальным результатам.

Функции потерь, оценивающие сходство таблиц, играют ключевую роль в генерации синтетических данных, так как именно они позволяют зафиксировать различия между реальными и синтетическими таблицами. Главная цель – добиться максимального сходства синтетических таблиц с реальными данными. Но существующие на сегодняшний день функции потерь обладают рядом недостатков, таких как неспособность адекватно оценивать качество таблиц, высокая вычислительная сложность, неудобство использования в обучении нейронных сетей и т.д.

В третьей главе диссертационной работы «**Алгоритмы генерации синтетических данных**» представлены математическая модель новой функции потерь, новый алгоритм генерации синтетических данных на основе GAN-сети, алгоритм генерации табличных данных, состоящих только из значений 0 и 1, с применением методов обучения с подкреплением.

Для оценки сходства таблиц разработана новая функция потерь, математическая модель которой включает следующие этапы:

Шаг 1. Вычисление суммы по строкам: пусть реальные данные заданы матрицей R, синтетические – матрицей S. Тогда x_{nm} – элемент -строки и m-столбца матрицы R, y_{nm} – элемент i-строки и j-столбца матрицы S. В данном случае размерность матриц: 766 строк и 9 столбцов. Суммируются элементы n-строки матрицы R и i-строки матрицы S:

$$r_n = \sum_m^9 x_{nm} \qquad (n=1,2,\ldots,9) \qquad (1)$$

$$s_i = \sum_j^9 y_{ij} \qquad (i=1,2,\ldots,9) \qquad (2)$$

Указанные выше суммы вычисляются для каждой n-строки и i-строки.

Шаг 2. Вычисление суммы по столбцам: вычисляются суммы элементов в m-столбце матрицы R и j-столбце матрицы S:

$$R_m = \sum_n^{766} x_{nm} \qquad (m=1,2,\ldots,766) \qquad (3)$$

$$S_j = \sum_i^{766} y_{ij} \qquad (j=1,2,\ldots,766) \qquad (4)$$

Шаг 3. Определение среднего значения вертикальных разностей: для каждого столбца разность между суммами элементов соответствующих столбцов матриц R и S берется по модулю, после чего вычисляется их среднее значение: $C=\{\left| R_m - S_j \right|\}$.

$$\mu = \frac{1}{9}\sum_{k=1}^{9} C_k \qquad (5)$$

Шаг 4. Для каждой строки вычисляется минимальное значение разности:

$$\delta_n = \min_i(|r_n - s_i|) \qquad (6)$$

В результате формируется множество M, состоящее из минимальных значений δ_n для каждой строки: $M = \{\delta_1, \delta_2, \ldots, \delta_n\}$, где *n= 1, 2, …766*.

Шаг 5. Вычисление процента повторяющихся минимальных значений: пусть N – количество минимальных значений, которые совпадают среди 766 строк. Тогда доля повторяющихся минимальных значений рассчитывается по формуле: P = N/766

Шаг 6. Определение среднего значения минимальных разностей:

$$\nu = \frac{1}{766}\sum_{n=1}^{766}\delta_n \quad (7)$$

Шаг 7. Вычисление горизонтальных потерь:

$$D = \frac{\nu}{P} \quad (8)$$

Шаг 8. Общая потеря для одного примера определяется следующим образом:

$$L_{bn} = D + \mu \quad (9)$$

Для каждого батча в процессе обучения получаем значения потерь L_{1bn}, L_{2bn} ,…, L_{kbn}, где k - количество примеров в наборе.

Шаг 9. Общая функция потерь вычисляется через квадратный корень из среднего значения потерь, полученных для всех батчей:

$$L_{TD} = \sqrt{\frac{1}{k}\sum_{l=1}^{k} L_l} \quad (10)$$

Алгоритм обучения с подкреплением (изображен на рис. 2) включает в себя следующие этапы:

Шаг 1. Ввод статистических данных в алгоритм.

Шаг 2. Выработка первичной базы данных на основе этих статистических данных и соответствующих формул.

Шаг 3. Перемешивание данных по столбцам с использованием алгоритма Фишера-Ятса.

Шаг 4. Перемешивание данных по столбцам с применением предложенного нового алгоритма.

Шаг 5. Ввод начальных значений коэффициентов P и B, а также параметров, управляющих их изменением: β_1, β_2, ε.

Шаг 6. Дополнительное перемешивание данных по заданному диапазону с использованием алгоритма Фишера-Ятса и значений коэффициентов P и B, адаптируя выборку под обучающую задачу

Шаг 7. Передача перемешанной обучающей выборки в нейронную сеть и ее обучение в течение 5 эпох.

Шаг 8. Тестирование нейронной сети, предварительно обученной на синтетической выборке, на реальной обучающей выборке.

Шаг 9. Сохранение синтетической обучающей выборки, если достигнут приемлемый уровень точности модели.

Традиционная GAN-сеть состоит из двух частей: генератора и дискриминатора. Эти две нейронные сети обучаются совместно в конкурентном взаимодействии. Генераторная сеть принимает случайный шум в качестве входных данных и преобразует его в синтетические образцы

данных. Эффективность генератора оценивается по его способности генерировать высококачественные, не повторяющиеся образцы, которые способны обмануть дискриминатор. Генератор достигает высокой способности к генерации путем минимизации функции потерь. Задача дискриминатора – оценивать вероятность того, что представленный образец является реальным. В процессе обучения дискриминатор учится отличать реальные данные, взятые из обучающего набора, от данных, искусственно сгенерированных генератором. В результате конкуренции между генератором и дискриминатором общая GAN-сеть начинает генерировать синтетические данные, максимально приближенные к реальным.

Рис. 2. Алгоритм генерации базы синтетических данных на основе заданных статистических данных

Для адаптации традиционной модели GAN к поставленной цели в нее внесены следующие три модификации: во-первых, на вход генератора подается не один, а два входных потока, вместо случайного шума используются статистические данные; во-вторых, для обучения генератора используется новая функция потерь, специально разработанная для оценки качества синтетических табличных данных; в-третьих, статистические данные внедряются во внутренние слои дискриминатора – не на вход, а за несколько слоев до выходного слоя. В данной модифицированной GAN-сети генератор включает такие слои, как ConvTranspose, BatchNormalization, Dropout, ReLU,

Flatten, Output, а дискриминатор включает такие слои, как Convolution, BatchNormalization, Dropout, ReLU, Flatten, Output.

Для генерации синтетических данных с помощью GAN-сети разработана модифицированная модель традиционной GAN-сети (см. рис. 3).

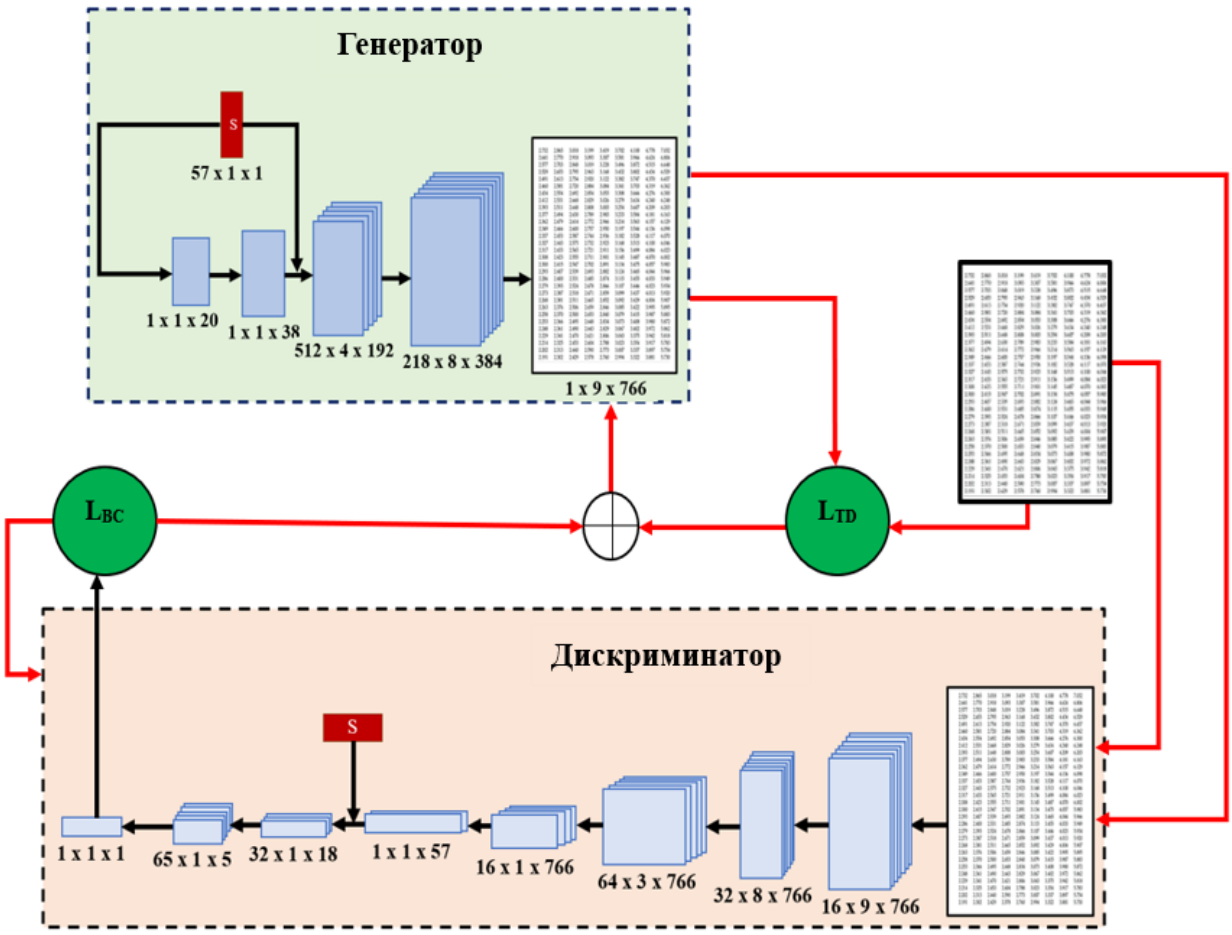


Рис. 3. Архитектура предложенной GAN-сети

Соответствующая функция потерь для дискриминатора формулируется следующим образом:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[lgD\left(x^{(i)}\right) + \lg(1 - D(G(z^{(i)}))) \right] \qquad (11)$$

Для генератора же:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} lg(1 - D(G(z^{(i)}))) + L_{TS} \qquad (12)$$

При этом, L_{TS} – дополнительная функция потерь.

GAN-сеть обучается генерации таблиц на основе статистических данных, поэтому для ее обучения требуется двойная база данных, состоящая из базы табличных данных и базы статистических данных, которая содержит статистические характеристики, соответствующие этим таблицам. Последовательность выполнения данного процесса представлена на рис. 4.

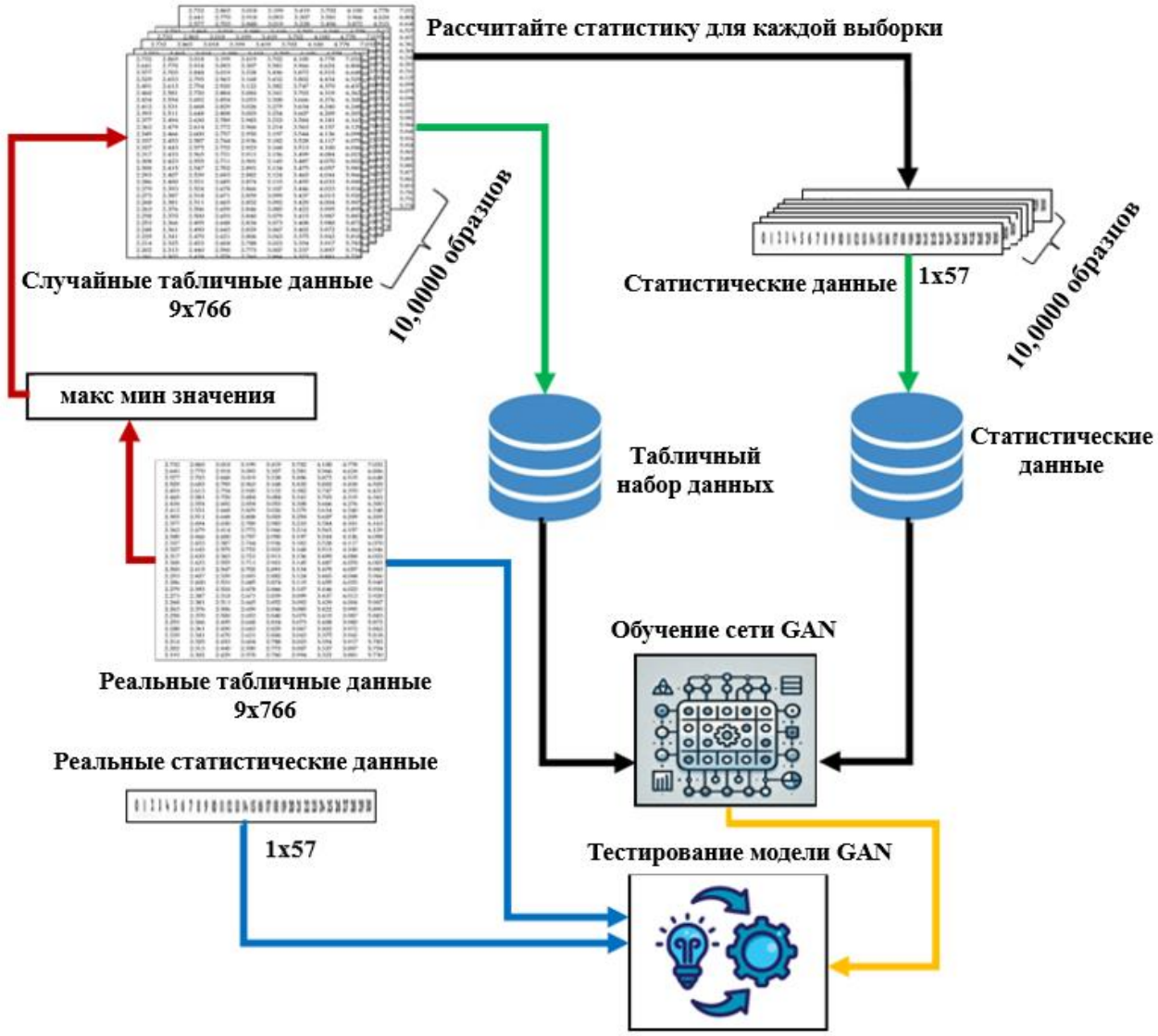


Рис. 4. Алгоритм генерации синтетической обучающей выборки с использованием GAN-сети

Шаг 1. Генерация базы статистических данных.

Шаг 2. Для каждой записи базы статистических данных генерируются первичные табличные данные. Формируется база табличных данных на основе объединения этих табличных данных.

Шаг 3. Обучение GAN-сети с использованием двух баз данных (табличной и статистической). В процессе обучения сопоставляются взаимосвязи между переменными в синтетической и реальной обучающей выборке.

Шаг 4. Оценка синтетических обучающих выборок, сгенерированных обученной GAN-сетью на основе реальных статистических данных, по критериям полезности и сходства.

Шаг 5. В случае удовлетворительного результата оценки сохраняются гиперпараметры архитектуры GAN-сети.

Четвертая глава диссертации «**Оценка синтетических обучающих выборок, сгенерированных с помощью алгоритмов искусственного интеллекта**» посвящена критериям оценки качества табличных

синтетических данных, их классификации и определениям. Сходство синтетических табличных данных, полученных с помощью новых моделей и алгоритмов, с реальными данными и их полезность для обучения алгоритмов искусственного интеллекта оценены на основе выбранных критериев. Оценка качества синтетических данных после их генерации является важнейшей задачей.

Таблица 1

Значения расстояния Вассерштейна

Навзвание атрибута	*Предложенный метод*	*GM*	*CTGAN*
Кол-во беременностей	0.077049	0.050503	0.090343
Уровень глюкозы	0.139552	0.029401	0.211049
Артериальное давление	0.119209	0.077486	0.135793
Толщина кожного покрова	0.166620	0.082580	0.063013
Инсулин	0.047145	0.381628	0.161786
MVN	0.155688	0.026149	0.111652
Генетические аспекты диабета	0.176867	0.051625	0.057591
Возраст	0.086909	0.022935	0.031314

В четвертой главе диссертации «**Оценка синтетических обучающих выборок, сгенерированных с помощью алгоритмов искусственного интеллекта**» приведены критерии оценки качества синтетических табличных данных, их классификация и определения. С использованием разработанных новых моделей и алгоритмов проведена оценка степени схожести синтетических табличных обучающих выборок с реальными, а также их полезности для обучения алгоритмов искусственного интеллекта на основе выбранных критериев. Оценка качества синтетических данных после их генерации является одной из важнейших задач. В настоящем исследовании для комплексной оценки синтетических табличных обучающих выборок (СОУВ) выделены следующие методы:

1. Методы, выбранные для оценки сходства синтетической табличной обучающей выборки: t-критерий Стьюдента, χ^2-критерий, расстояние Вассерштейна

2. Методы, выбранные для оценки полезности: точность, F1-score, специфичность, чувствительность и различия между значениями, полученными по тестам TRTR и TSTR.

Для оценки сходства использовалось расстояние Вассерштейна. При этом расстояние Вассерштейна рассчитывалось отдельно для каждого атрибута обучающей выборки. Результаты расчетов представлены в таблице 1. Из таблицы видно, что расстояние Вассерштейна для всех атрибутов в выборках, полученных с помощью предложенного метода, меньше критического значения (0,3). Это указывает на выполнение условия схожести.

Таблица 2

Результаты теста TSTR

	Критерии оценки			
Название метода	*Точность (%)*	*Чувствительность (%)*	*Специфичность (%)*	*F1-оценка (%)*
Реальная обучающая выборка	0.7532	0.7554	0.7532	0.7542
SDV	0.6494	**0.7648**	0.6494	0.6494
CTGAN	0.3312	0.388	0.3312	0.2812
Предложенный метод	**0.7078**	0.6966	**0.7078**	**0.6926**

Для оценки полезности использовался тест TSTR. В качестве модели ИИ применялась модель Random Forest (RF). Для каждой модели рассчитаны показатели: точность, специфичность, чувствительность и F1-оценка. Информация о результатах TSTR-теста приводится в таблице 2.

ЗАКЛЮЧЕНИЕ

На основе проведтнных исследований по теме диссертации «Модели и алгоритмы для генерации синтетических обучающих выборок на основе статистических данных» сформулированы следующие выводы:

1. Проведен сравнительный анализ актуальных задач и проблем, связанных с генерацией синтетических табличных обучающих выборок, а также методов получения синтетических медицинских данных. В результате анализа обоснована высокая потребность в табличных синтетических обучающих выборках, предназначенных для прогнозирования риска различных хронических заболеваний.

2. Классифицированы функции потерь, используемые при генерации синтетических обучающих выборок с помощью GAN-сетей, и выполнен их сравнительный анализ. Установлено, что существующие методы оценки сходства таблиц слабо адаптированы к табличным данным, в связи с чем выявлена необходимость разработки специальной функции потерь для оценки близости табличных данных.

3. В соответствии с вышеуказанным требованием разработан алгоритм генерации полноценных синтетических медицинских данных на основе только статистической информации с применением нейронных сетей и метода обучения с подкреплением.

4. Определено, что использование только одной функции потерь при генерации табличных синтетических обучающих выборок на основе GAN-сети дает ограниченные результаты. Добавление дополнительной функции потерь к основной позволяет повысить эффективность работы сети.

5. При генерации табличных синтетических медицинских обучающих выборок с использованием GAN-сети разработана новая функция потерь,

которую можно применять в процессе обучения GAN- и VAE-сетей. Данная функция потерь позволяет повысить качество генерируемых таблиц.

6. Разработан алгоритм генерации синтетических табличных данных с использованием GAN-сети. Этот алгоритм обеспечивает возможность генерации необходимых табличных обучающих выборок на основе статистических данных.

7. На основе статистических обучающих выборок разработана обучающая выборка, пригодная для обучения алгоритмов искусственного интеллекта с целью прогнозирования риска развития сахарного диабета II типа за 5 лет до возможного заболевания. Использование этой выборки при обучении нейронной сети повысило эффективность прогноза риска развития диабета II типа на 5 %.

8. Выявлено, что синтетические обучающие выборки, полученные с помощью модифицированной GAN-сети, отличаются высокой достоверностью, сходством с реальными данными и полезностью для обучения моделей искусственного интеллекта, что подтверждено результатами оценки по различным критериям.

SCIENTIFIC COUNCIL AWARDING SCIENTIFIC DEGREES DSc.13/05.05.2023.T.07.03 AT TASHKENT UNIVERSITY OF INFORMATION TECHNOLOGIES

TASHKENT UNIVERSITY OF INFORMATION TECHNOLOGIES

NASIMOVA NIGORAKHON MIZROBOVNA

MODELS AND ALGORITHMS OF GENERATING THE SYNTHETIC DATASET BASED ON STATISTICAL DATA

05.01.11 – Digital technologies and artificial intelligence

ABSTRACT OF THE DISSERTATION OF DOCTOR OF PHILOSOPHY (PHD) IN TECHNICAL SCIENCES

Tashkent – 2025

Тема диссертации доктора философии (PhD) по техническим наукам зарегистрирована в Высшей аттестационной комиссии при Министерство высшего образования, науки и инноваций Республики Узбекистан за B2025.1.PhD/T5320.

Диссертация выполнена в Ташкентском университете информационных технологий.

Автореферат диссертации на трех языках (узбекский, английский, русский (резюме)) размещен на веб странице (www.tuit.uz) и на Информационно-образовательном портале «Ziyonet» (www.ziyonet.uz).

Научный руководитель:	**Муминов Баходир Болтаевич** доктор технических наук, профессор
Официальные оппоненты:	**Мухамедиева Дилдора Кабиловна** доктор технических наук, доцент
	Атаджанов Ибрагим Равшанбекович доктор технических наук
Ведущая организация:	**Национальный исследовательский университет «Ташкентский институт инженеров ирригации и механизации сельского хозяйства»**

Защита диссертации состоится «25» июня 2025 г. в 14⁰⁰ часов на заседании научного совета DSc.13/05.05.2023.T.07.03 при Ташкентском университете информационных технологий. (Адрес: 100084, г. Ташкент, ул. Амира Темура, 108. Тел.: (99871) 238-64-43; e-mail: ilmiy_kengash@tuit.uz).

С диссертацией можно ознакомиться в Информационно-ресурсном центре Ташкентского университета информационных технологий (регистрационный номер №357). (Адрес: 100084, г. Ташкент, ул. Амира Темура, 108. Тел.: (99871) 238-64-70).

Автореферат диссертации разослан «12» июня 2025 года.
(протокол рассылки №9 от «12» июня 2025 г.).

М.М.Камилов
Председатель Научного совета по присуждению учёных степеней, доктор технических наук, академик АН РУз

Н.А.Эгамбердиев
Ученый секретарь Научного совета по присуждению учёных степеней, доктор философии по техническим наукам

Н.О.Рахимов
Председатель научного семинара при Научном совете по присуждению учёных степеней, доктор технических наук, доцент

INTRODUCTION (abstract of PhD dissertation)

The aim of the research work is to develop models and algorithms for generating synthetic dataset used to train chronic disease prediction algorithms based on artificial intelligence algorithms and statistical data.

The object of the research work is statistical data on chronic diseases, GAN network, loss function, reinforcement learning algorithm.

The scientific novelty is as follows:

a mathematical model of the new loss function for training GAN networks has been developed, which evaluates the similarity of synthetic tabular dataset to real dataset, taking into account their statistical characteristics;

a mathematical model of shuffling the values in table columns consisting of “0” and “1” has been developed based on the binary partial inversion method;

an algorithm for generating synthetic tabular dataset based on reinforcement learning by shuffling variables has been developed;

the input and output parts of the GAN network were modified, a newly developed loss function was used in addition to the loss function used to train it, and a model and algorithm for generating synthetic training samples were developed based on this modified architecture.

Implementation of the research results. Based on the developed research methods and algorithms, the research results were put into practice in the following areas:

a computer program that allows diagnosing type 2 diabetes based on 8 symptoms, created using a synthetic training sample based on the algorithm "GAN algorithm for generating synthetic tabular data based on medical statistical data", was successfully implemented at the Multidisciplinary Central Polyclinic of the Olot District Medical Association. This program helps to increase the accuracy of diagnosis of diseases associated with diabetes mellitus, and its implementation into practice has been assessed as effective (certificate of the Ministry of Digital Technologies dated February 25, 2025, No. 33-8/1255). The implemented system increased the productivity of industry specialists by 5%;

a computer program that allows for heart attack detection based on synthetic training samples generated based on the "Medical Synthetic Data Generation Algorithm" has been successfully implemented in the Margilan interdistrict perinatal center. This program helps to increase the accuracy of diagnosis of diseases associated with diabetes mellitus, and its implementation into practice has been recognized as effective (certificate of the Ministry of Digital Technologies dated February 25, 2025, No. 33-8/1255). The implemented system increased the productivity of specialists by 7%;

The applications "Algorithm for generating synthetic medical data" and "GAN algorithm for generating synthetic tabular medical data based on statistical information" were implemented at the Center for Advanced Training of Medical Workers (reference of the Ministry of Digital Technologies dated February 25, 2025 No. 33-8/1255). These algorithms, based on the developed algorithms and software solutions, played an important role in the educational process in the field of

information technology and artificial intelligence. They allow students to work with training samples, use them to train algorithms, identify errors and understand how synthetic samples can serve as an alternative to real ones in cases of their absence, as well as understand their advantages and limitations, which helps to increase the efficiency of the educational process.

Publication of the research results. The total number of scientific works on the subject of the research is 28, including 7 journal papers (5 papers in international journals and 2 papers in republican journals) recommended in scientific publication of the Supreme Attestation Commission of the Republic of Uzbekistan are published and 5 certificates of registration of the software created for the computer are received.

The structure and the volume of the dissertation. The dissertation consists of an introduction, five chapters, conclusion, references, abbreviations and appendixes. The dissertation consists of 118 pages.

E’LON QILINGAN ISHLAR RO‘YXATI
СПИСОК ОПУБЛИКОВАННЫХ РАБОТ
LIST OF PUBLISHED WORKS

I bo‘lim (I часть; I part)

1. R.H. Nasimov, B. Muminov, N.Nasimova, S. Mirzahalilov, “A New Approach to Classifying Myocardial Infarction and Cardiomyopathy Using Deep Learning // International Conference on “Information Science and Communications Technologies (ICISCT) 2020”. Tashkent -2020. p. 1-5 (05.00.00; 30.10.2020 №287/9 – son rayosat qarori)
2. B.Muminov,R.Nasimov, N.Nasimova, K. Abdurashidova, M. Abdullaev, Comparative Analysis of the Results of Algorithms for Dilated Cardiomyopathy and Hypertrophic Cardiomyopathy Using Deep Learning // International Conference on “Information Science and Communications Technologies (ICISCT) 2021.” Tashkent -2021. p. 1-5 (05.00.00; 30.10.2021 №525-son rayosat qarori.)
3. Б.Б. Мўминов, Р.Ҳ. Насимов, Н.М. Насимова, “Сурункали касалликларни масофадан мониторинг қилувчи дастурий иловасини ишлаб чиқиш”, ТАТУ хабарлари, 3,4(63-64), 2022, 23-32 б. (05.00.00; № 31)
4. Abdusalomov, A.B., Nasimov, R., Nasimova, N., Muminov, B. Whangbo, T.K. Evaluating Synthetic Medical Images Using Artificial Intelligence with the GAN Algorithm. Sensors 2023, *23*, 3440. p.1-21 (Web of Science, №1)
5. Z. Tagmatova, A. Abdusalomov, R. Nasimov, A. H. Dogru, Y.I. Cho, “New Approach for Generating Synthetic Medical Data to Predict Type 2 Diabetes”, “Bioengineering” 10, no.9:1031, 2023, p. 1-14 (Web of Science, №1)
6. Rashid Nasimov, Nigorakhon Nasimova, Sanjar Mirzakhalilov², Gul Tokdemir, Mohammad Rizwan, Akmalbek Abdusalomov, Young-Im Cho: “GAN-Based Novel Approach for Generating Synthetic Medical Tabular Data”, Bioengineering, 11, no 12:1288. 2024, p.1-17 (Web of Science, №1)
7. Nasimov Rashid, Nasimova Nigorakhon, Muminov Bakhodir, “Novel approach of generating fully sinthetic medical tabular dataset”, Digital transformation and Artificial intelligence: problems, Innovations and trends, 1st international scientific - practical conference, Sep 11, Tashkent 2024, p. 238-243 (05.00.00; OAK Rayosatining 2024-yil 30-iyuldagi 358/3-son qarori)

II bo‘lim (II часть; II part)

8. Б.Б.Мўминов, Н.М.Насимова, Р.Ҳ.Насимов, “Сунъий интеллектдан фойдаланиб синтетик рақамли тиббий тасвирларни ҳосил қилиш ва баҳолаш”.// Журнал “Иқтисодиёт ва инновацион технологиялар”, № 3/2022. Toshkent -2022. 324-338 б.

9. R.Nasimov, N.Nasimova, 2-tur diabet kasalligini bashoratlash algoritmlarining qiyosiy tahlili, “Raqamli iqtisodiyot va axborot texnologiyalari” ilmiy elektron jurnali. № 4(12). Toshkent -2023. 71-76 b.

10. R.H. Nasimov, B.B. Muminov, N.Nasimova, S. Mirzahalilov “Algorithm of Automatic Differentiation of Myocardial Infarction from Cardiomyopathy based on Electrocardiogram” International Conference on “Application of Information and Communication Technologies (AICT) 2020”, Tashkent -2020. P. 219-223

11. B.B.Mo‘minov, N.M.Nasimova, “Gibrid neyron tarmoqlari yordamida sintetik tibbiy tasvirlarni hosil qilish algoritmi”, “Иқтисодиёт тармоқларининг инновацион ривожланишида ахборот-коммуникация технологияларининг аҳамияти” мавзуидаги Республика илмий –техник анжуманининг маърузалар тўплами, 2 -қисм, Тошкент. *2022, 493-495 б.*

12. B. B. Mo‘minov, R. H. Nasimov, N. M. Nasimova”Xronik kasalliklarni masofadan monitoring qilish ilovalarini baholash protokollari tahlili”, International Scientific and Technical Conference “Digital Technologies: Problems and Solutions for Practical Implementation in an Industry” April 27-28, 2022. P. 36-38

13. N. Nasimova “Semi-Empiric Future Importance Dependency Evolution Method to Increase Diagnosis Accuracy of Chronic Disease”, “Innovative Development in The Global Science”, International Scientific and Technical Conference, Boston, USA, 2022, 96-100 b.

14. R. Nasimov, N. Nasimova, K. Botirjon, M. Abdullayev, “Deep learning algorithm for classifying dilated cardiomyopathy and hypertrophy cardiomyopathy in transport workers”, “Internet of things, Smart Spaces and Next generation Networks and Systems”, NEW2AN-2022, Lecture Notes in Computer Science, Vol 13772, Springer Cham. p. 218-230

15. R. Nasimov, N. Nasimova, B. Mo‘minov, “Hybrid Method For Evaluating Feature Importance for Predicting Chronic Heart Diseases”, “Applications, Trends and Opportunities” International Conference on Information Science and Communications Technologies ICISCT 2022, 28-30 September 2022

16. R. H. Nasimov, N. M. Nasimova “The importance of a mobile application for monitoring chronic diseases in the study of AI methods for medical personnel”, Информатизация образования и методика электронного обучения: цифровые технологии в образовании” Материали VI Международной научной конференции Красноярск, 20-23 сентября 2022 г., Част 3, 455-458 б.

17. R. H. Nasimov, N. M. Nasimova “Xronik kasalliklarni monitoring qilishda ko‘maklashuvchi dasturlar tahlili”, “Kompyuter ilmlari va muhandislik Texnologiyalari” mavzusidagi Xalqaro miqyosidagi ilmiy-texnik anjuman materiallari to‘plami, 2-qism, 14-15 oktyabr 2022, 217-219 b.

18. R.H.Nasimov, N.M.Nasimova “Surunkali kasalliklarni monitoring qilishda grafik usullardan samarali foydalanishning ahamiyati”, Internation conference on “Recent advances in intellegent information and communication technologies-ISPC 2022”, Tashkent. 2022, p. 88-91

19. R. H. Nasimov, N. M.Nasimova “Surunkali kasalliklarni kunlik monitoring qiluvchi mobil ilovani ishlab chiqish”, “Matematik modellashtirish, axborot-

kommunikatsiya texnologiyalarning dolzarb masalalari” Respublika ilmiy-texnik anjumanining ma’ruzalar to‘plami, Nukus, 2022, 258-259 b.

20. N.Nasimova, R.Nasimov, F.Muhiddinova, “Bemorlarning tibbiy ma’lumotlari xavfsizligi ta’minlash uchun sintetik ma’lumotlar bazasi hosil qilish usullari tahlili”, “Zamonaviy axborotlarni kriptografik himoyalash vositalarini amaliyotga tadbiq qilish” mavzusidagi Respublika onlayn ilmiy-amaliy konferensiyasi ma’ruzalar to‘plami,Toshkent, 2023. 148-153 b,

21. R.Nasimov, N.Nasimova, F.Muhiddinova, Bemorlarning tibbiy ma’lumotlari xavfsizligi ta’minlash uchun sintetik ma’lumotlar bazasi hosil qilish usullari tahlili // “Zamonaviy axborotlarni kriptografik himoyalash vositalarini amaliyotga tatbiq qilish” Respublika onlayn ilmiy-amaliy anjumani ma’ruzalar to‘plami, Toshkent -2023. 148-152 b.

22. M.Raximov, R.Nasimov, N.Nasimova, A.Usmanxodjaeva, Surunkali kasalliklarni masofadan monitoring qilishning shifokor ilovasi // O‘zbekiston Respublikasi Adliya vazirligi. Elektron hisoblash mashinalari uchun yaratilgan dasturning rasmiy ro‘yxatdan o‘tkazilganligi to‘g‘risidagi guvohnomasi, DGU 22557, 27.02.2023

23. M.Raximov, R.Nasimov, N.Nasimova, A.Usmanxodjaeva, Surunkali kasalliklarni masofadan monitoring qilishning bemor ilovasi // O‘zbekiston Respublikasi Adliya vazirligi. Elektron hisoblash mashinalari uchun yaratilgan dasturning rasmiy ro‘yxatdan o‘tkazilganligi to‘g‘risidagi guvohnomasi, DGU 22282, 17.02.2023

24. M.Raximov, M. Ochilov, R.Nasimov, N.Nasimova, A.Usmanxodjaeva, Surunkali kasalliklarni masofadan monitoring qilishning bemor ilovasi // O‘zbekiston Respublikasi Adliya vazirligi. Elektron hisoblash mashinalari uchun yaratilgan dasturning rasmiy ro‘yxatdan o‘tkazilganligi to‘g’‘risidagi guvohnomasi, DGU 22555, 27.02.2023

25. R. Nasimov, N. Nasimova, B. Muminov, A. Usmanxodjayeva, G.Sobirova, A. Abdusalomov, “Development of Fully Synthetic Medical Database Shuffling Method”, “Internet of Things, Smart Spaces, and Next Generation Networks and Systems”, Springer. NEW2AN ruSMART; Lecture Notes in Computer Science. 2023, p. 1-10

26. N.Nasimova, R.Nasimov,“To‘la sintetik tibbiy ma’lumotlarni ishlab chiqishning O‘zbekiston uchun amaliy ahamiyati va zarurati tahlili” “Sun’iy intellekt nazariyasi va amaliyoti: tajriba, muammolar va istiqbollar” Respublika ilmiy-texnik anjumanining ma’ruzalar to‘plami, 2-qism, Toshkent, 2024, 350-353b.

27. N.Nasimova, R.Nasimov “Tibbiy sintetik ma’lumotlarni hosil qilish algoritmi”, O‘zbekiston Respublikasi Adliya vazirligi, Elektron hisoblash mashinalari uchun yaratilgan dasturning rasmiy ro‘yxatdan o‘tkazilganligi to‘g’risidagi Guvohnoma. DGU 39735, 30.05.2024

28. N.Nasimova, B.Mo‘minov, R.Nasimov “Tibbiy statistik ma’lumotlar asosida sintetik jadvalli ma’lumotlarni hosil qilishning GAN algoritmi, O‘zbekiston Respublikasi Adliya vazirligi, EHM uchun yaratilgan dasturning rasmiy ro‘yxatdan o‘tkazilganligi to‘g‘risidagi Guvohnoma. DGU 43742, 09.11.2024

Avtoreferat “Al-Xorazmiy avlodlari” O‘zbekiston ilmiy jurnali taxririyatida taxrirdan o‘tkazildi hamda o‘zbek, rus va ingliz tillaridagi matnlari mosligi tekshirildi.