Шинжлэх Ухаан, Технологийн Их Сургууль
Мэдээлэл, Холбооны Технологийн Сургууль

Чулуунбаатарын Амгалан-Очир

Дижитал орчин дахь хуурамч видео, зураг илрүүлэх систем хөгжүүлэх

Бакалаврын төгсөлтийн ажил

Улаанбаатар хот

Шинжлэх Ухаан, Технологийн Их Сургууль
Мэдээлэл, Холбооны Технологийн Сургууль

Кибер аюулгүй байдлын тэнхим

Дижитал орчин дахь хуурамч видео, зураг илрүүлэх систем хөгжүүлэх

Мэргэжлийн индекс: D061202000000002306
Мэргэжил: Кибер аюулгүй байдал

Удирдагч: Магистр Х.Уянгаа
Зөвлөгч: доктор (Ph.D), Ч.Эрдэнэбат
Доктор(Ph.D) дэд профессор Я.Дашдорж
Гүйцэтгэгч: Ч.Амгалан-Очир

Улаанбаатар хот
2026 он 6 сар

Батлав. Кибер аюулгүй байдлын тэнхимийн эрхлэгч:

..... /доктор (Ph.D), дэд профессор Б.Мөнхбаяр/

Удирдагч: ..... /Магистр Х.Уянгаа/

Дипломын төсөл гүйцэтгэх төлөвлөгөө

Дипломын төслийн сэдэв:
Монгол: " Дижитал орчин дахь хуурамч видео, зураг илрүүлэх систем хөгжүүлэх"
Англи: " Development of Fake Images and Videos Detection in Digital Environments"

Төслийн зорилго: Дижитал орчин дахь хуурамч зураг, видеог илрүүлэх програм хангамжийг нээлттэй сан ашиглан хөгжүүлэх.

Гүйцэтгэх оюутны овог нэр: Ч.Амгалан-Очир/B221870067/
Холбоо барих утас: 80248078, 89509967

№	Ажлын бүлэг, хэсгийн нэр	эзлэх хувь	дуусах хугацаа
Бүлэг №1. Хуурамч зураг, бичлэгийн технологийн суурь судалгаа			
1	1.1 Хуурамч контентын ойлголт, төрөл ба хэрэглээ 1.2 Хуурамч контент үүсгэх технологиуд 1.3 Аюулгүй байдлын эрсдэл, хор уршиг ба чиг хандлага 1.4 Бүлгийн дүгнэлт	20%	III.28
Бүлэг №2. Хуурамч контент илрүүлэлт ба баталгаажуулалтын арга, хэрэгсэл, алгоритмын харьцуулсан шинжилгээ			
2	2.1 Хуурамч зураг , видео илрүүлэлтийн аргачлал ба нээлттэй эхийн хэрэгсэл, алгоритмын харьцуулалт 2.2 Өгөгдөл бэлтгэл ба үнэлгээний аргууд 2.3 Контентын баталгаажуулалт ба дижитал мөр 2.4 Бүлгийн дүгнэлт	40%	IV.21
Бүлэг №3. Техникийн хэрэгжилт ба интеграцчилал			
3	3.1 Илрүүлэгчийн загвар хэрэгжүүлэлт 3.2 Хуурамч зураг, бичлэг илрүүлэх програмын хөгжүүлэлт 3.3 Хэрэгжүүлэлтийн үр дүн 3.3 Бүлгийн дүгнэлт	40%	V.18
Бүлэг №4. Ерөнхий дүгнэлт			

Төлөвлөгөөг боловсруулсан оюутан: ..... /Ч.Амгалан-Очир/

ТӨГСӨЛТИЙН АЖЛЫН ҮЗЛЭГИЙН ХУУДАС

Оюутны код: B221870067
Оюутны нэр: Ч.Амгалан-Очир
Сэдвийн монгол нэр: ” Дижитал орчин дахь хуурамч видео, зураг илрүүлэх систем хөгжүүлэх”
Сэдвийн англи нэр: ” Development of Fake Images and Videos Detection in Digital Environments”
Удирдагч багш: Магистр Х.Уянгаа
Зөвлөгч багш: доктор (Ph.D), Ч.Эрдэнэбат, Доктор(Ph.D) дэд профессор Я.Дашдорж

№	Үзлэгийн гүйцэтгэл	Гүйцэтгэлийн 30% -с багагүй байна.	Огноо	Удирдагч Магистр Х.Уянгаа багшийн гарын үсэг
1	Үзлэг-1		III/01-III/06	

Багшийн товч зөвлөгөө, тайлбар:

.....
.....
.....
.....
.....
.....
.....

Үзлэг-1 хийсэн багш: ..... /Магистр Х.Уянгаа/

№	Үзлэгийн гүйцэтгэл	Авсан оноо (10 оноо)	Гүйцэтгэлийн 50% -с багагүй байна.	Огноо	доктор (Ph.D), Ч.Эрдэнэбат багшийн гарын үсэг
1	Үзлэг-2			IV/15-V/01	

Багшийн товч зөвлөгөө, тайлбар:

.....
.....
.....
.....
.....
.....
.....

Үзлэг-2 хийсэн багш: ..... /доктор (Ph.D), Ч.Эрдэнэбат/

ТӨГСӨЛТИЙН АЖЛЫН ҮЗЛЭГИЙН ХУУДАС

Оюутны код: B221870067
Оюутны нэр: Ч.Амгалан-Очир
Сэдвийн монгол нэр: ” Дижитал орчин дахь хуурамч видео, зураг илрүүлэх систем хөгжүүлэх”
Сэдвийн англи нэр: ” Development of Fake Images and Videos Detection in Digital Environments”
Удирдагч багш: Магистр Х.Уянгаа
Зөвлөгч багш: доктор (Ph.D), Ч.Эрдэнэбат, Доктор(Ph.D) дэд профессор Я.Дашдорж

№	Үзлэгийн гүйцэтгэл	Авсан оноо (10 оноо)	Гүйцэтгэлийн 70% -с багагүй байна.	Огноо	Доктор(Ph.D) дэд профессор Я.Дашдорж багшийн гарын үсэг
1	Үзлэг-3			V/04-V/15	

Багшийн товч зөвлөгөө, тайлбар:

.....
.....
.....
.....
.....
.....
.....

Үзлэг-3 хийсэн багш: ..... /Доктор(Ph.D) дэд профессор Я.Дашдорж/

№	Үзлэгийн гүйцэтгэл	Гүйцэтгэлийн 90% -с багагүй байна.	Огноо	Удирдагч Магистр Х.Уянгаа багшийн гарын үсэг
1	Үзлэг-4		V/13-V/17	

№	Удирдагч Магистр Х.Уянгаа багшийн үнэлгээ (30 оноо)	Огноо	Удирдагч багшийн гарын үсэг
1		V/17	

Удирдагч багш: ..... /Магистр Х.Уянгаа/

Жич: Удирдагч багш өөрийн үнэлгээгээ 30 хүртэл оноогоор өгөх ба үнэлгээ тавьсан хуудсыг оюутанд буцааж өгөлгүй төгсөлтийн нарийн бичгийн даргад хураалгана уу.

ТӨГСӨЛТИЙН АЖЛЫН ЯВЦ

№	Хийж гүйцэтгэсэн ажил	Биелсэн хугацаа	Удирдагчийн гарын үсэг
1	Бүлэг №1. Хуурамч зураг, бичлэгийн технологийн суурь судалгаа	2026-03-28	
2	Бүлэг №2. Хуурамч контент илрүүлэлт ба баталгаажуулалтын арга, хэрэгсэл, алгоритмын харьцуулсан шинжилгээ	2026-04-21	
3	Бүлэг №3. Техникийн хэрэгжилт ба интеграцчилал	2026-05-18	
4	Бүлэг №4. Ерөнхий дүгнэлт	2026-05-25	

Ажлын товч дүгнэлт

.....

.....

.....

.....

.....

.....

.....

Удирдагч: ..... /Магистр Х.Уянгаа/

ЗӨВШӨӨРӨЛ

Оюутан Ч.Амгалан-Очир–н бичсэн төгсөлтийн ажлыг УШК-д хамгаалуулахаар тодорхойлов.

Салбарын эрхлэгч: ..... /доктор (Ph.D), дэд профессор Б.Мөнхбаяр/

Шинжлэх Ухаан, Технологийн Их Сургууль
Мэдээлэл, Холбооны Технологийн Сургууль

ШҮҮМЖИЙН ХУУДАС

Кибер аюулгүй байдлын тэнхимийн төгсөх курсийн оюутан Ч.Амгалан-Очир-н "Дижитал орчин дахь хуурамч видео, зураг илрүүлэх систем хөгжүүлэх" сэдэвт төгсөлтийн ажлын шүүмж.

1. Төслөөр дэвшүүлсэн асуудал, үүнтэй холбоотой онолын материал уншиж судалсан байдал. Энэ талаар хүмүүсийн хийсэн судалгаа, түүний үр дүнг уншиж тусгасан эсэх. (6 оноо)

   .....
   .....
   .....
   .....
   .....
   .....
   .....

2. Төслийн ерөнхий агуулга. Шийдсэн зүйлүүд, хүрсэн үр дүн. Өөрийн санааг гарган, харьцуулалт хийн, дүгнэж байгаа чадвар. (6 оноо)

   .....
   .....
   .....
   .....
   .....
   .....
   .....

3. Эмх цэгцтэй, стандарт хангасан өөрөөр хэлбэл диплом бичих шаардлагуудыг биелүүлсэн эсэх. Төсөлд анзаарагдсан алдаанууд, зөв бичгийн болон өгүүлбэр зүйн гэх мэт /Хуудас дугаарлагдаагүй, зураг хүснэгтийн дугаар болон тайлбар байхгүй, шрифт хольсон, хувилсан зүйл ихээр оруулсан/. (6 оноо)

   .....
   .....
   .....
   .....
   .....

4. Төслөөр орхигдуулсан болон дутуу болсон зүйлүүд. Цаашид анхаарах хэрэгтэй зүйлүүд. (6 оноо)

   .....
   .....
   .....
   .....
   .....
   .....
   .....

5. Төслийн талаар онцолж тэмдэглэх зүйлүүд. (6 оноо)

   .....
   .....
   .....
   .....
   .....
   .....
   .....

6. Ерөнхий оноо. (30 оноо)

   .....

Шүүмж бичсэн: ..... /доктор (Ph.D), Д.Бямбадорж/

Ажлын газар: .....

Хаяг (Утас) .....

Зохиогчийн эрх хамгаалал

Миний бие Ч.Амгалан-Очир, "Дижитал орчин дахь хуурамч видео, зураг илрүүлэх систем хөгжүүлэх" сэдэвт энэ ажил нь минийх бөгөөд дараахыг нотолж байна. Үүнд:

- Горилогч энэ ажлыг тус сургуулиас боловсролын зэрэг авахаар бүхэлд нь буюу голлон хийсэн болно.
- Энэ ажлын аль нэг хэсгийг тус сургуульд эсвэл өөр байгууллагад боловсролын зэрэг, мэргэшил авахаар өмнө нь илгээсэн бол түүнийгээ тодорхой заасан болно.
- Бусад хүмүүсийн хэвлүүлсэн ажлаас зөвлөгөө авсан бол түүнийгээ үндэслэсэн болно.
- Бусад хүмүүсийн ажлаас ишлэл хийхдээ эх үүсвэрийг нь заасан болно.
- Миний ажилд тусалсан голлох бүх эх үүсвэрт талархаж байна.
- Ажлыг бусадтай хамтарсан бол алийг нь бусад хүмүүс хийсэн болохыг тодорхой заасан болно.

Гарын үсэг: _____

Огноо: _____

Шинжлэх Ухаан, Технологийн Их Сургууль
Мэдээлэл, Холбооны Технологийн Сургууль

Хураангуй

Дижитал орчин дахь хуурамч видео, зураг илрүүлэх систем хөгжүүлэх

Ч.Амгалан-Очир
B221870067@must.edu.com

Түлхүүр үгс: *дипфейк, хуурамч зураг, хуурамч видео, хиймэл зураг, нүүрэн дипфейк, ансамбль илрүүлэлт, контент баталгаажуулалт*

Дижитал орчинд зураг, видео контентын бодитой байдал нь мэдээллийн аюулгүй байдал, олон нийтийн итгэлцэл, хувь хүний нэр хүнд болон байгууллагын баталгаажуулалттай шууд холбоотой асуудал болж байна. Сүүлийн жилүүдэд дипфейк, хиймэл зураг үүсгэх, царай солих, уруулын хөдөлгөөн тааруулах болон бодит мэт дүрс бүтээх технологиуд хурдтай хөгжсөнөөр хуурамч контентыг энгийн хэрэглэгч ялган танихад улам хүндрэлтэй болсон. Иймээс зураг, видео агуулгыг автоматаар шинжлэх, хуурамч байх магадлалыг тооцоолох, үр дүнг тайлбарлах хэрэгцээ нэмэгдэж байна.

Энэхүү дипломын ажлын зорилго нь дижитал орчин дахь хуурамч зураг, видеог илрүүлэх боломжтой программ хангамж боловсруулах, мөн уг асуудлын онолын суурь, өгөгдөл бэлтгэл, илрүүлэх аргачлал болон үнэлгээний үндсийг нэгтгэн судлахад оршино. Судалгааны хүрээнд хуурамч контентын төрөл, үүсгэх технологи, тухайлбал GAN, авто-кодлогч, диффузийн загвар, трансформерт суурилсан үүсгэгч загварууд болон тэдгээрээс үлдэх дүрсний бүтэц, давтамжийн ул мөр, сэргээн босголтын ялгаа, фрейм хоорондын уялдааг авч үзсэн.

Техникийн хэрэгжилтийн түвшинд зураг болон видео шинжлэх desktop application боловсруулсан. Зураг шинжилгээнд CLIP-д суурилсан илрүүлэгч, DIRE, CNN Detection, UniversalFakeDetect зэрэг аргуудыг ашиглаж, видео шинжилгээнд Xception, EfficientNet-B4, F3Net, SPSL болон MN-FaceDF загварууудыг нэгтгэсэн. Систем нь хэрэглэгчийн интерфэйс, файл оруулах хэсэг, метаданс унших, нүүр илрүүлэх, фрейм сонгох, илрүүлэгч загваруудын оноо нэгтгэх, Hive AI API холболт, шинжилгээний түүх хадгалах боломжуудтайгаар хэрэгжсэн.

Судалгааны нэг чухал хэсэг болгон Монгол хүний царайны онцлогийг тусгах MN-ImageDF өгөгдлийн санг бүрдүүлсэн. Wikipedia болон Wikimedia Commons-оос Монгол хүмүүстэй холбоотой 459 хүний мэдээлэл цуглуулж, MTCNN алгоритмаар 410 нүүрийг тогтвортой илрүүлсэн. Үүний дараа бодит болон хуурамч ангилалд хамаарах нийт 1124 ширхэг 224×224 хэмжээтэй нүүрний сгор бэлтгэж, сургалт, баталгаажуулалт, тестийн өгөгдөл болгон ашигласан.

Монгол царайны өгөгдөл дээр ResNet-18 архитектурт суурилсан MN-FaceDF загварыг сургаж, 168 тестийн дээж дээр үнэлгээ хийсэн. Туршилтын үр дүнд Accuracy = 0.845, Balanced Accuracy = 0.818, Precision = 0.865, Recall = 0.703, F1-score = 0.776, Macro F1 = 0.829, ROC AUC = 0.865 үзүүлэлт гарсан. Энэ үр дүн нь бүрдүүлсэн Монгол царайны өгөгдөл болон хэрэгжүүлсэн загвар нь бодит ба хуурамч нүүрний зургийг ялгах боломжтойг харуулсан.

Талархал

Энэхүү дипломын ажлыг бичихэд туслалцаа үзүүлсэн удирдагч багш Х.Уянгаа болон зөвлөх багш Ч.Эрдэнэбат , дэд профессор Я.Дашдорж ,гэр бүл ,ШУТИС-ийн Мэдээлэл холбоо технологийн сургуулийн Кибер аюулгүй байдал тэнхимийн багш нарт талархсанаа илэрхийлье.

Удиртгал

Сүүлийн жилүүдэд хиймэл оюун ухаан, гүн сургалт, дүрс болон дуу боловсруулах технологи хөгжсөнөөр зураг, видео, дуу хоолойг бодит мэтээр өөрчлөх, шинээр үүсгэх боломж нэмэгдсэн. Үүний нэг хэлбэр болох дипфейк технологи нь хүний нүүр царай, хөдөлгөөн, уруулын хөдөлгөөнийг ашиглан бодит мэт хуурамч контент үүсгэдэг. Энэхүү дипломын ажилд дараах асуудлыг авч үзэв.

1. Хуурамч зураг, видео, дуу хоолойн үндсэн ойлголт, хэрэглээг судлах.
2. GAN, авто-кодлогч, диффузийн загвар, трансформер болон дуу хоолойн дипфейк үүсгэх технологийг тайлбарлах.
3. Хуурамч контентоос үүсэх хувь хүн, байгууллага, нийгэмд нөлөөлөх аюулгүй байдлын эрсдэлийг тодорхойлох.
4. Хуурамч зураг, видео илрүүлэх аргачлал, нээлттэй эхийн хэрэгсэл, өгөгдлийн сан, үнэлгээний хэмжүүрийг харьцуулан судлах.
5. Контентын баталгаажуулалт, дижитал мөр, гарал үүслийн мэдээлэл, таних тэмдэглэгээний нөхцөлүүдийг авч үзэх.
6. Хуурамч зураг, видео илрүүлэх программын ерөнхий бүтэц, хэрэгжүүлэх боломжийг тодорхойлох.

Сэдвийг сонгох үндэслэл

Дижитал орчинд зураг, видео нь мэдээлэл дамжуулах хамгийн хурдан бөгөөд нөлөөтэй хэлбэрүүдийн нэг болсон. Гэвч хиймэл оюун ухаанд суурилсан үүсгэгч технологи хөгжсөнөөр бодит мэт харагдах хуурамч зураг, видео үүсгэх боломж нэмэгдэж байна. Энэ нь мэдээллийн бодит байдал, хувь хүний нэр төр, байгууллагын итгэлцэл болон цахим аюулгүй байдалд бодитой эрсдэл бий болгож байна [1]. Энэ эрсдэл нь зөвхөн онолын түвшинд бус, бодит цахим хэрэглээтэй шууд холбоотой юм. DataReportal-ийн 2026 оны тайланд Монгол Улсад 2025 оны эцсийн байдлаар 2.93 сая интернэт хэрэглэгч байсан бөгөөд интернэтийн нэвтрэлт нийт хүн амын 83.0%-д хүрсэн байна. Мөн 2.70 сая сошиал медиа хэрэглэгчийн бүртгэлтэй байсан нь нийт хүн амын 76.5%-тай тэнцэж, нийт интернэт хэрэглэгчдийн 92.2% нь дор хаяж нэг сошиал платформ ашиглаж байгааг харуулж байна [2]. Энэ нь Монголын цахим орчинд зураг, бичлэг зэрэг медиа контентууд маш хурдан тархах нөхцөл бүрдсэнийг илтгэнэ.

1. Монгол Улсад интернэт болон сошиал медиа хэрэглээ өндөр түвшинд хүрсэн тул видео, зурагууд богино хугацаанд олон хүнд хүрэх боломжтой болсон.
2. Хуурамч зураг, видео нь хувь хүний нэр төрд халдах, байгууллагын нэр хүндийг унагах, олон нийтийн итгэлцлийг сулруулах, худал мэдээлэл түгээх эрсдэлтэй.

3. Хүний нүүр царай, дүр төрхийг зөвшөөрөлгүй ашиглан залилан хийх, нийгмийн инженерчлэл явуулах эрсдэлүүд нэмэгдэж байна [1].

4. Хуурамч зураг, видео илрүүлэх асуудалыг бүрэн илрүүлэх программ хангамж дутмаг байна. Ялангуяа зураг, видео шахагдах, дахин кодлогдох, чанар буурах, гэрэлтүүлэг өөрчлөгдөх үед илрүүлэлтийн үр дүн тогтворгүй болох боломжтой.

5. Орчин үеийн дипфейк контент нь нүүрний хэлбэр, арьсны бүтэц, гэрэлтүүлэг, сүүдэр, уруулын хөдөлгөөн, кадр хоорондын уялдаа зэрэг олон төрлийн дүрсний шинжийг ашигладаг. Иймээс зураг болон видео өгөгдөлд суурилсан илрүүлэлтийн арга шаардлагатай.

6. Монгол хүний дүр төрх, нүүрний онцлог, орчны гэрэлтүүлэг, камерын чанар зэрэг нөхцөл нь олон улсын өгөгдлийн санд хангалттай тусгагдаагүй. Иймээс Монгол Улсын орчинд тохирсон өгөгдөл бэлтгэх, туршилт хийх, программын түвшинд шалгах шаардлагатай.

7. Ийм нөхцөлд Монголын нөхцөл байдалд ашиглах боломжтой, зураг болон видео хуурамч контентыг илрүүлэх программын шийдэл боловсруулах нь мэдээллийн аюулгүй байдлын түвшинд бодит ач холбогдолтой.

Иймээс энэхүү сэдэв нь хиймэл оюун ухаанаар үүсгэсэн хуурамч зураг, видеоны эрсдэлийг судлах, түүнийг илрүүлэх аргачлалыг турших, мөн Монголын цахим орчинд тохирсон программын шийдэл боловсруулах хэрэгцээ шаардлагад үндэслэн сонгосон болно.

Судалгааны шинэлэг тал

Өмнөх судалгааны ажлуудад хуурамч зураг, видео илрүүлэх асуудлыг олон талаас нь авч үзсэн байна. Rössler нарын FaceForensics++ ажил нь DeepFakes, Face2Face, FaceSwap, NeuralTextures зэрэг нүүр хувиргах аргуудаар үүсгэсэн өгөгдөл дээр илрүүлэгч загваруудыг үнэлэх benchmark өгөгдлийн сан бүрдүүлсэн [3]. Li нарын Celeb-DF судалгаа нь өмнөх өгөгдлийн сангуудын дүрсний чанар бодит цахим орчинд тархдаг дипфейк видеотой төдийлөн ойр биш байсныг онцолж, илүү бодит мэт өндөр чанартай дипфейк видео өгөгдөл санал болгосон [4]. Wang нарын CNNDetection ажил нь нэг төрлийн үүсгэгч загвар дээр сургасан ангилагч бусад харагдаагүй үүсгэгч загварын зурагт тодорхой хэмжээнд ижил шинжтэй болохыг харуулсан [5]. Мөн DIRE арга нь диффуз загвараар үүсгэсэн зургийг сэргээн засах алдаанд тулгуурлан ялгах боломжийг судалсан бол GenImage өгөгдлийн сан нь олон төрлийн орчин үеийн үүсгэгч загвар дээр хиймэл зураг илрүүлэгчийг шалгах нөхцөл бүрдүүлсэн [6, 7]. Эдгээр ажлууд нь хуурамч контент илрүүлэх судалгааны суурийг бүрдүүлсэн боловч ихэнх нь олон улсын өгөгдлийн сан, лабораторийн benchmark орчин болон тусгай загварын үнэлгээнд төвлөрсөн байдаг. Иймээс энэхүү дипломын ажлын давуу ба шинэлэг тал нь өмнөх судалгааны ажилуудад хэрэглэгдсэн өгөгдөл, илрүүлэлтийн арга, үнэлгээний хэмжүүрийг нэгтгэн авч үзэж, зураг болон видео хуурамч контентыг программын түвшинд шинжлэх шийдэл болгон хэрэгжүүлэхэд оршино.

1. Өмнөх судалгаануудад ашиглагдсан өгөгдлийн сан, илрүүлэгч загвар, benchmark үнэлгээний аргачлалыг нэгтгэн судалсан.

2. Хуурамч зураг болон видео илрүүлэлтийг тусад нь авч үзэхээс илүүтэй, нэг программын ажиллагаанд холбон авч үзсэн.

3. Зургийн түвшний оноо, видео кадрын түвшний оноо, эцсийн ангиллын үр дүнг нэгтгэн тооцох зарчмыг тодорхойлсон.

4. Олон улсын өгөгдлийн сангууд Монгол орчны нүүр царай, гэрэлтүүлэг, камерын чанар, цахим хэрэглээний нөхцөлийг бүрэн төлөөлөхгүй байж болох тул Монгол орчинд тохирсон өгөгдөл бэлтгэх хэрэгцээг тусгасан.

5. Илрүүлэлтийн үр дүнг зөвхөн “жинхэнэ” эсвэл “хуурамч” гэсэн дүгнэлтээр хязгаарлахгүй, магадлалын оноо, итгэлийн түвшин, ашигласан загварын мэдээлэлтэй хамт харуулах байдлаар программын шийдэл боловсруулсан.

Зорилго, зорилтууд

Энэхүү дипломын ажлын зорилго нь дижитал орчин дахь хуурамч зураг, видеог нээлттэй сан ашиглаж, хөгжүүлэх

1. Хуурамч контент, хиймэл медиа, дипфейк гэсэн ойлголтыг ялган тодорхойлох.

2. Хуурамч контентоос үүсэх аюулгүй байдлын эрсдэл, хор уршгийг тодорхойлох.

3. Илрүүлэлтийн үр дүнг үнэлэх хэмжүүр, өгөгдөл бэлтгэх аргачлалыг авч үзэх.

4. Контентын баталгаажуулалт, дижитал мөр, таних тэмдэгийн тайлбарлах.

5. Хуурамч зураг, видео илрүүлэх программын бүтэц, модулийг зохион байгуулалт зэргийг хэрэгжүүлэх боломжийг тодорхойлох.

Товчилсон үгс

AI	**A**rtificial **I**ntelligence
ML	**M**achine **L**earning
DL	**D**eep **L**earning
GAN	**G**enerative **A**dversarial **N**etwork
CNN	**C**onvolutional **N**eural **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork
LSTM	**L**ong **S**hort-**T**erm **M**emory
GRU	**G**ated **R**ecurrent **U**nit
ViT	**Vi**sion **T**ransformer
CLIP	**C**ontrastive **L**anguage–**I**mage **P**re-training
TTS	**T**ext-**t**o-**S**peech
VC	**V**oice **C**onversion
ASV	**A**utomatic **S**peaker **V**erification
MFCC	**M**el-**F**requency **C**epstral **C**oefficients
CQCC	**C**onstant **Q** **C**epstral **C**oefficients
FFT	**F**ast **F**ourier **T**ransform
ROC	**R**eceiver **O**perating **C**haracteristic
AUC	**A**rea **U**nder the **C**urve
EER	**E**qual **E**rror **R**ate
GUI	**G**raphical **U**ser **I**nterface
PDF	**P**ortable **D**ocument **F**ormat
JSON	**J**ava**S**cript **O**bject **N**otation
CSV	**C**omma-**S**eparated **V**alues

Гарчиг

Зургийн жагсаалт

Хүснэгтийн жагсаалт

БҮЛЭГ 1

Хуурамч зураг, бичлэгийн технологийн суурь судалгаа

1.1 Хуурамч контентын ойлголт, төрөл ба хэрэглээ

Сүүлийн жилүүдэд хиймэл оюун ухаан, дүрс боловсруулалт, гүн сургалтын аргачлал эрчимтэй хөгжсөнөөр бодит мэт харагдах зураг, видео болон хүний нүүрний хөдөлгөөнийг хиймлээр үүсгэх боломж нэмэгдсэн. Үүний үр дүнд **хуурамч контент** нь зөвхөн зураг засварлах энгийн ойлголтоор хязгаарлагдахгүй, мэдээллийн үнэн зөв байдал, цахим орчны ёс зүй, хувь хүний нэр хүнд, байгууллагын итгэлцэл болон мэдээллийн аюулгүй байдалтай шууд холбоотой асуудал болж байна.

Монгол ардын “Мянга сонсохоор нэг үз” гэх зүйр үг нь хүн аливаа мэдээллийг нүдээр харсан тохиолдолд илүү бодитой хүлээн авах хандлагатайг илэрхийлдэг. Гэвч өнөөгийн дижитал орчинд хиймэл оюун ухаанаар бодит мэт зураг, видео үүсгэх боломж өргөжсөнөөр харагдаж буй дүрс бүрийг үнэн гэж үзэх боломжгүй болсон. Иймээс зураг, видео контентын үнэн зөв байдлыг шалгах, хуурамч эсэхийг илрүүлэх асуудал нь мэдээллийн аюулгүй байдал болон олон нийтийн итгэлцлийг хамгаалах чухал судалгааны чиглэл болж байна.

Судалгааны бүтээлүүдэд **хуурамч контент**, **хиймэл контент**, **дипфейк** гэсэн нэр томьёо өргөн хэрэглэгддэг. **Хуурамч контент** нь бодит байдлыг гажуудуулах, хүнийг төөрөгдүүлэх зорилгоор тухайн контентийн эзнийг өөрчлөх эсвэл зохиомлоор бүтээсэн агуулгыг илэрхийлнэ. **Хиймэл контент** нь тооцооллын загвар, үүсгэгч алгоритм болон компьютерийн боловсруулалтын аргаар бүхэлд нь эсвэл хэсэгчлэн үүсгэсэн зураг, видео зэрэг дижитал агуулгыг хамарсан өргөн ойлголт юм. Харин **дипфейк** нь гүн сургалт, нейрон сүлжээ, үүсгэгч загварт тулгуурлан хүний дүр төрх, нүүрний хувирал, хөдөлгөөнийг бодит мэтээр дуурайлган бүтээсэн хиймэл видео контентын дэд ангилалд хамаарна [8].

Зураг 1.1: Зураг, бичлэг болон олон модаль хуурамч контентын ерөнхий ангилал

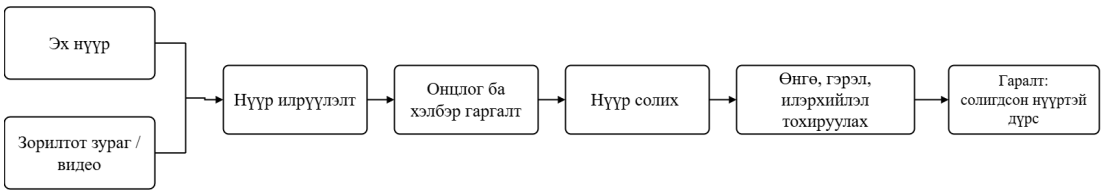
Зураг 1.1-д дипфейк буюу хиймэл бичлэг, зурагт суурилсан, видеонд суурилсан болон олон модалийг харуулав [8, 9].

Царай солих

Царай солих нь нэг хүний нүүр царайг өөр хүний зураг, бичлэг дээр байршуулж, тухайн хүнийг өөр үйлдэл хийж байгаа мэт харагдуулах арга юм. Энэ төрлийн өөрчлөлт нь дипфейкийн сонгодог хэлбэрүүдийн нэг бөгөөд авто-кодлогчид суурилсан анхны дипфейк системүүдээс эхлээд орчин үеийн өндөр нарийвчлалтай царай солих систем хүртэл тасралтгүй хөгжиж ирсэн.

Царай солих үед зөвхөн нүүрийг давхарлаж тавихад хангалтгүй. Нүүрний хэлбэр, арьсны өнгө, гэрэл сүүдэр, нүүрний өнцөг, нүд амны байрлал, нүүрний хувирал, хөдөлгөөний дараалал зэрэг олон хүчин зүйлийг хооронд нь уялдуулах шаардлагатай. Хэрэв эдгээр зүйлс сайн таарвал хуурамч дүрс нь хүний нүдэнд ялгагдахгүй хэмжээнд хүрдэг. Хэрэв нийлүүлэлт алдаатай үед нүүрний хэлбэр, арьсны бүтэц, гэрэлтүүлгийн зөрчил, нүдний орчмын үл нийцэл зэрэг мөр үлдэх боломжтой [8]

Зураг 1.2-д царай солих үйл явцын үндсэн дарааллыг схемээр харуулсан.



Зураг 1.2: Царай өөрчлөх процесс

Уруулын хөдөлгөөнийг тааруулах Уруулын хөдөлгөөнийг тааруулах гэдэг нь бичлэг дээрх хүний ам, уруул болон нүүрний доод хэсгийн хөдөлгөөнийг өөр ярианы дуу авиа, үг хэллэгтэй нийцүүлэн өөрчлөх аргыг хэлнэ. Энэ арга нь зөвхөн нүүр солихоос ялгаатай бөгөөд ярианы авиа, амны хэлбэр, нүүрний булчингийн хөдөлгөөн, цаг хугацааны уялдаа зэрэг динамик шинжүүдийг нарийн тааруулах шаардлагатай байдаг [8, 9].

Уруулын хөдөлгөөнийг тааруулах үед авиа, үе, үг, ам нээх ба хаах хөдөлгөөнөөс гадна эрүү, хацар болон нүүрний ерөнхий хувирал хоорондоо нийцтэй байх нь чухал. Хэрэв зөвхөн уруулын хэсгийг хөдөлгөж, бусад нүүрний хөдөлгөөнтэй уялдуулж чадахгүй бол тухайн бичлэг хуурамч болох нь амархан мэдэгддэг.

Харин орчин үеийн уруулын хөдөлгөөн тааруулах загварууд нь ярианы дуунаас амны хэлбэр, хөдөлгөөний дарааллыг сурч, өгөгдсөн царай дээр илүү бодитой хөдөлгөөн үүсгэх боломжтой болсон. Иймээс энэ төрлийн технологи нь дуу-визуал дипфейкийг бодит мэт харагдуулах нэг гол хүчин зүйл болж байна [9].

Зураг 1.3-т бодит видео цуглуулах, нүүр хувиргалт хийх болон CNN загвараар жинхэнэ/хуурамч гэж ангилах ерөнхий схемийг [3]-ын судалгааны ажлаас иш татан харуулав.

Зураг 1.3: Бодит видео цуглуулах, нүүр хувиргалт хийх процесс

Хуурамч контентын эерэг ба сөрөг хэрэглээ

Хуурамч контент үүсгэх технологи нь дан ганц сөрөг утгатай биш юм. Зөв зохистой хэрэглээнд кино урлаг, тоглоом, дубляж, виртуал дүр, боловсролын үзүүлэн, түүхэн дүрслэл сэргээх, олон хэлний контентыг илүү хүртээмжтэй болгох зэрэг салбарт бодит үнэ цэн бүтээж болно [10, 11].

Гэвч энэ технологийн сөрөг хэрэглээ нийгэм, байгууллага, хувь хүнд ноцтой эрсдэл үүсгэдэг. Хуурамч зураг, бичлэг, дуу нь худал мэдээлэл тараах, хүний нэр хүндэд халдах, байгууллагын нэрээр хуурамч мэдэгдэл гаргах, санхүүгийн шилжүүлэг хийлгэх, хэрэглэгчийг төөрөгдүүлэн шийдвэр гаргуулах, цахим дээрэлхэлт үйлдэх, улс төрийн зорилготой хуурамч яриа түгээх зэрэг олон төрлийн хор уршигтай [8, 12].

Зураг 1.4: Dalí Lives төслийн бодит хэрэглээний жишээ [13].

Эерэг хэрэглээний бодит жишээ болгон The Dalí Museum-ийн Dalí Lives төслийг авч үзэж болно. Тус төсөл нь хиймэл оюун ухаан ашиглан Сальвадор Далиг музейн дэлгэцээр амьд мэт харуулж, үзэгчидтэй интерактив байдлаар харилцах боломж бүрдүүлсэн байна. Үүнийг Зураг 1.4-д харуулав.

1.2 Хуурамч контент үүсгэх технологиуд

Хуурамч зураг, бичлэг үүсгэх технологийн гол зорилго нь бодит дүрсний хэв шинжийг сурч, түүнтэй төстэй шинэ дүрс бий болгох, эсвэл байгаа дүрсийг бодит мэтээр өөрчлөхөд оршдог бөгөөд эдгээр ойлголтууд онолын суурь болж байна [8, 11, 14, 15].

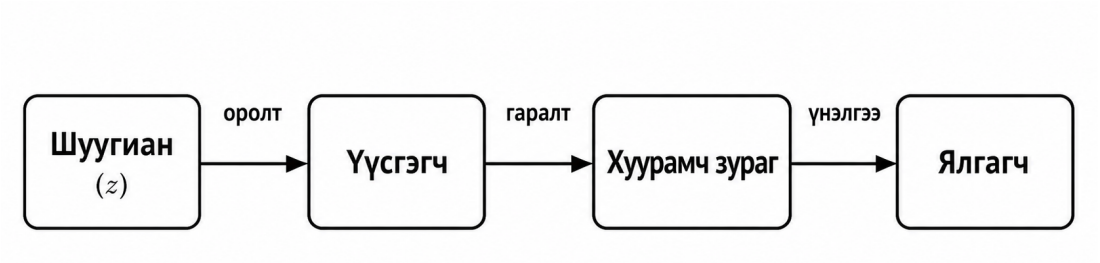
1. **GAN** — Хоёр сүлжээ хоорондоо өрсөлдөн суралцаж, бодит мэт зураг, видео үүсгэдэг загвар.
2. **Авто-кодлогчид суурилсан дипфейк** — Нүүрний ерөнхий байрлал, хөдөлгөөнийг хадгалж, танигдах шинжийг өөр хүний нүүрээр сольдог арга.
3. **Диффузийн загвар** — Дуу чимээнээс дүрсийг шат дараатай сэргээх замаар өндөр чанартай синтетик зураг үүсгэдэг загвар.

4. **Трансформерт суурилсан үүсгэгч загвар** — Текст, зураг, видео зэрэг мэдээллийн хамаарлыг сурч, өгөгдсөн зааврын дагуу хиймэл контент үүсгэдэг загвар.

GAN загварын онолын үндэс

GAN буюу Generative Adversarial Network нь үүсгэгч ба ялгагч гэсэн хоёр хэсгээс бүрддэг. Үүсгэгч нь бодит мэт харагдах дүрс гаргахыг оролддог бол ялгагч нь тухайн дүрс жинхэнэ эсвэл хиймэл эсэхийг ялгадаг. Энэ хоёр хэсэг хоорондоо өрсөлдөн суралцдаг бүтэц нь GAN-ийн гол онцлог бөгөөд үүсгэгдсэн дүрсийн чанар аажмаар сайжрах нөхцөл болдог [14].

GAN загварын үндсэн санаа нь бодит өгөгдлийн тархалтыг шууд дүрэмчилж бичихгүйгээр өгөгдлөөс өөрөөс нь суралцан шинэ дүрс үүсгэхэд оршино. Үүсгэгч нь ялгагчийг хуурахуйц бодит мэт дүрс гаргахыг хичээдэг бол ялгагч нь бодит болон хиймэл дүрсийг улам сайн ялгахыг зорьдог. Ийнхүү хоёр сүлжээний харилцан өрсөлдөөнөөс үүдэн үүсгэгчийн гаргаж буй дүрс аажмаар бодит мэт болж сайжирдаг.



Зураг 1.5: GAN загварын үндсэн бүтэц ба хуурамч зураг үүсгэх ерөнхий үйл явц

Зураг 1.5-д GAN буюу өрсөлдөөнт үүсгэгч сүлжээний үндсэн ажиллах зарчмыг хялбаршуулан үзүүлэв.

GAN загвар нь нүүр царай, хөрөг зураг, хиймэл дүрс бүтээхэд ихээхэн нөлөө үзүүлсэн. Гэвч үүсгэсэн зурагт зарим тохиолдолд ирмэгийн гажиг, нүүрний жижиг хэсгүүдийн үл нийцэл, гэрэлтүүлгийн зөрүү, давтамжийн хиймэл шинж зэрэгт үлдэх боломжтой [8, 14].

$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{x\sim p_{data}}\left[\log D(x)\right] + \mathbb{E}_{z\sim p_z}\left[\log\left(1 - D(G(z))\right)\right] \tag{1.1}$$

Энд G нь үүсгэгч, D нь ялгагч, x нь бодит өгөгдөл, z нь санамсаргүй оролт юм. Үүсгэгч нь ялгагчийг хуурахуйц дүрс үүсгэхийг зорьдог бол ялгагч нь бодит болон хиймэл дүрсийг зөв ялгахыг сурдаг.

Авто-кодлогчид суурилсан дипфейкийн онол

Дипфейкийн практик хэрэглээнд өргөн тархсан эхний архитектурын нэг нь **авто-кодлогчид суурилсан дипфейк** юм. Авто-кодлогч нь дүрсийг кодлогчоор далд орон зай руу шахаж, дараа нь тайлагчаар дахин сэргээх зарчимтай. Нүүр солих нөхцөлд хоёр өөр хүний нүүрийг нэг төрлийн дундын кодлогчоор далд төлөөлөл болгон хувиргаж, харин ялгаатай тайлагчуудаар сэргээх замаар нэг хүний нүүрний танигдах байдлыг нөгөө хүний хөдөлгөөн, байрлалтай уялдуулах боломжтой болдог [8].

Авто-кодлогчид суурилсан царай солих аргын гол санаа нь хүний нүүрний байрлал, нүүрний хөдөлгөөн, илэрхийлэл, гэрэлтүүлэг зэрэг ерөнхий шинжийг хадгалж,

харин танигдах байдлыг өөр хүний нүүрний онцлог руу хувиргах явдал юм. Өөрөөр хэлбэл, тухайн хүн хэрхэн хөдөлж, ямар өнцгөөс харагдаж байгааг хадгалсан хэвээр нүүрний танигдах шинжийг өөрчилдөг.

Хэрэв загвар далд орон зай дотор танигдах байдал, илэрхийлэл, байрлал, гэрэлтүүлэг зэрэг шинжийг сайн салгаж сурвал царай солих илүү бодит болдог. Харин энэ салгалт сайн ажиллаагүй үед нүдний ирмэг, нүүрний зах, арьсны бүтэц, фрейм хоорондын тасалдах үзэгдэл зэрэг илрүүлэх боломжтой шинжүүд үлддэг [8].

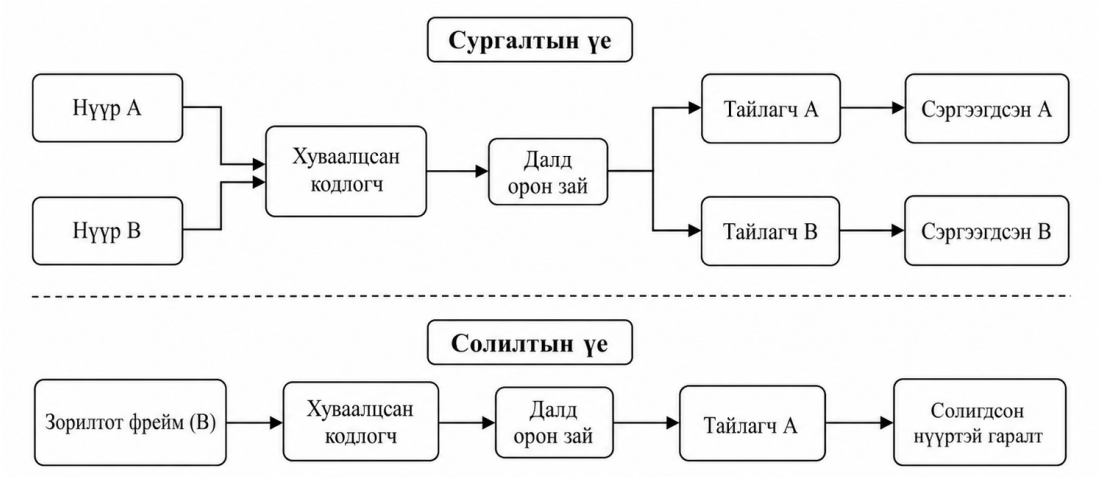
$$z = E(x),\ \hat{x} = D(z) \tag{1.2}$$

Энд E нь кодлогч, D нь тайлагч, z нь далд төлөөлөл, $\hat{x}$ нь сэргээгдсэн дүрс юм. Загварын зорилго нь эх дүрс болон сэргээгдсэн дүрсний ялгааг багасгах явдал юм.

$$L_{rec} = \|x - \hat{x}\|_2^2 \tag{1.3}$$

Царай солих үед зорилтот хүний нүүрний далд төлөөллийг нөгөө хүний тайлагчаар сэргээж, танигдах шинжийг өөрчилдөг.

$$\hat{x}_{B\to A} = D_A\left(E(x_B)\right) \tag{1.4}$$



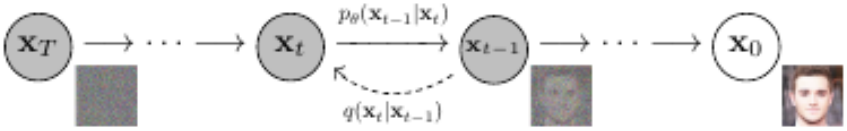
Зураг 1.6: Авто-кодлогчид суурилсан царай солих дипфейкийн ажиллах зарчим

Зураг 1.6-д авто-кодлогчид суурилсан царай солих аргын үндсэн зарчмыг харуулав. Сургалтын үед хоёр хүний нүүрийг хуваалцсан кодлогчоор далд орон зайд хувиргаж, тус бүрийн тайлагчаар сэргээдэг. Харин солилтын үед зорилтот хүний байрлал, хөдөлгөөн, илэрхийллийг хадгалан нөгөө хүний танигдах шинжийг үүсгэдэг.

Диффузийн загварын онолын үндэс

Диффузийн загвар нь сүүлийн үеийн хиймэл зураг үүсгэх хамгийн хүчтэй аргуудын нэг болсон. Энэ аргын үндсэн санаа нь бодит зурагт эхлээд бага багаар шуугиан нэмж, эцэст нь танигдахгүй болтол нь сарниулах явдал юм. Дараа нь уг шуугианыг алхам алхмаар арилгаж, дүрсийг буцаан сэргээх үйл явцыг загвар сурдаг [15].

Диффузийн загварын сургалтын үед эх өгөгдөлд үе шаттайгаар шуугиан нэмэгдэж, дараа нь загвар энэ процессийн эсрэг чиглэлийг буюу шуугианаас дүрс сэргээх чадварыг эзэмшдэг. Ийм зарчим нь өндөр чанартай, нарийн бүтэцтэй, гэрэл зурагтай төстэй хиймэл зураг үүсгэх боломжийг бүрдүүлдэг.



Зураг 1.7: Диффузийн загварын шуугиан нэмэх ба сэргээх ерөнхий зарчим

Зураг 1.7-д диффузийн загварын үндсэн зарчмыг харуулав. Уг арга нь бодит дүрсэнд шуугиан нэмээд, дараа нь шуугианыг шат дараатай арилгах замаар бодит мэт дүрс сэргээдэг.

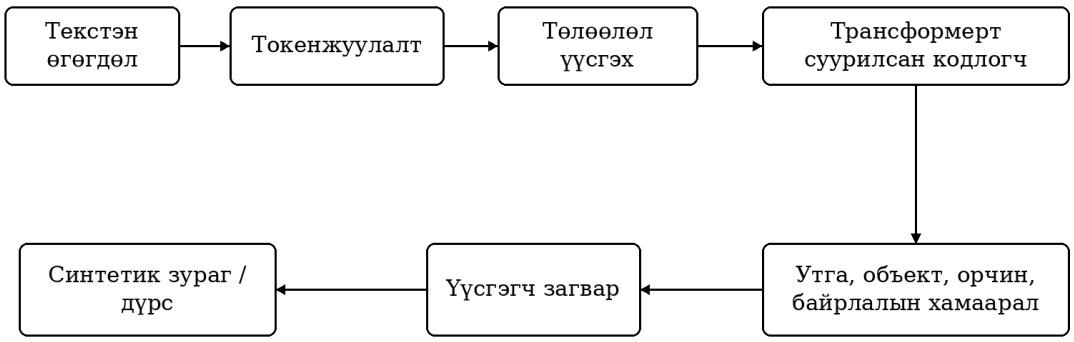
$$\mathrm{x_t} = \sqrt{\bar{\alpha}_\mathrm{t}}\mathrm{x}_0 + \sqrt{1-\bar{\alpha}_\mathrm{t}}\epsilon \tag{1.5}$$

Энд x_0 нь эх дүрс, $\mathrm{x_t}$ нь шуугиантай дүрс, ϵ нь шуугиан, $\bar{\alpha}_\mathrm{t}$ нь тухайн үе шатны хадгалагдсан мэдээллийн хэмжээг илэрхийлнэ.

$$\mathrm{L_{DDPM}} = \mathbb{E}\left[\|\epsilon - \epsilon_\theta(\mathrm{x_t}, \mathrm{t})\|_2^2\right] \tag{1.6}$$

Трансформерт суурилсан үүсгэгч загвар

Трансформер загварын гол онцлог нь өөрт чиглэсэн анхаарлын механизм ашиглан өгөгдлийн элементүүдийн хоорондын хамаарлыг тооцоолох чадвартай байдаг. Энэ механизм нь өгөгдлийн хэсгүүдийн хоорондын холын хамаарлыг давталт ашиглахгүйгээр, нэгэн зэрэг боловсруулах боломж олгодог. [11].



Зураг 1.8: Трансформерт суурилсан загварын схем

Зураг 1.8-д трансформерт суурилсан загвар текстэн өгөгдлийг боловсруулж, үүсгэгч загварт дамжуулан синтетик зураг, дүрс үүсгэх ерөнхий зарчмыг харуулав.

Трансформер загварт өгөгдлийн хэсгүүдийн хоорондын хамаарлыг анхаарлын механизмаар тооцдог.

$$\mathrm{Attention(Q, K, V)} = \mathrm{softmax}\left(\frac{\mathrm{QK^T}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \tag{1.7}$$

Энд Q нь асуулга, K нь түлхүүр, V нь утга, $\mathrm{d_k}$ нь түлхүүрийн хэмжээ юм. Энэ томьёо нь текстийн үг, объект, байрлал, орчны мэдээллийн хоорондын хамаарлыг тооцоход хэрэглэгддэг.

Текстээс зураг үүсгэх ерөнхий хэлбэрийг дараах байдлаар тооцно.

$$\hat{\mathrm{x}} = \mathrm{Gen}\left(\mathrm{T}_{\mathrm{enc}}(\mathrm{s})\right) \tag{1.8}$$

Энд s нь текстэн өгөгдөл, $\mathrm{T}_{\mathrm{enc}}$ нь трансформерт суурилсан кодлогч, Gen нь үүсгэгч загвар юм.

Бодит цагийн дипфейк

Бодит цагийн дипфейк нь шууд харилцааны үед хүний дүр төрх, нүүрний хөдөлгөөн, дуу хоолойг бодит цагт өөрчлөх технологи юм. Видео дуудлага, шууд дамжуулалт, онлайн уулзалт зэрэг орчинд ашиглагдах боломжтой тул аюулгүй байдлын эрсдэл өндөртэй [10].

Энэ төрлийн системд дүрсний чанараас гадна боловсруулалтын хурд, хоцрогдол багатай байдал, аудио-визуал хөд, системийн тогтвортой ажиллагаа чухал байдаг. Нүүрний хөдөлгөөн, уруулын хөдөлгөөн, харц болон дуу хоолой зэрэг нь бодит цагтай уялдаатай өөрчлөгдвөл хуурамч эсэхийг танихад хүндрэлтэй болдог.

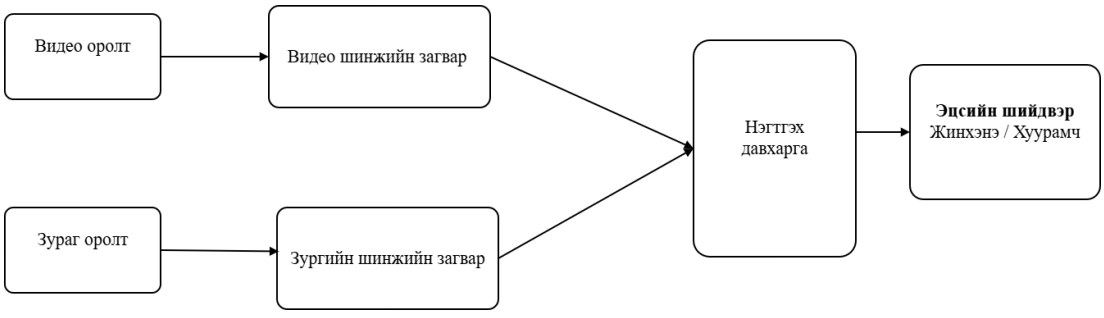
Иймээс бодит цагийн дипфейк нь дараах эрсдэл үүсгэж байна.

1. **Зайнаас ярилцлага хийх үед** тухайн хүний дүр төрх, дуу хоолойг дуурайн өөр хүн мэт оролцох боломжтой.
2. **Ажилд авах үйл явцад** үнэн бодит байдлыг шалгахад хүндрэл болох боломжтой.
3. **Онлайн баталгаажуулалтын үед** хүний нүүр, дуу хоолойгоор итгэл төрүүлж, хамгаалалтын шалгалтыг даван гарах эрсдэлтэй.
4. **Үйлчилгээний байгууллагын харилцаанд** байгууллага эсвэл хэрэглэгчийн дүрээр залилан мэхлэх нөхцөл бүрдүүлж болно.

Олон модаль хуурамч контент

Олон модаль хуурамч контент нь зураг, видео, дуу, текст зэрэг хэд хэдэн төрлийн өгөгдлийг нэгтгэн ашиглаж, илүү үнэмшилтэй хиймэл агуулга үүсгэхийг хэлнэ. Орчин үеийн дипфейк нь зөвхөн нүүр эсвэл дууг тусад нь дуурайлгахаас гадна нүүрний хөдөлгөөн, уруулын хөдөлгөөн, дуу хоолой, ярианы агуулгыг хамтад нь нийцүүлэх чиглэлд хөгжиж байна [8, 9].

Ийм төрлийн дипфейк нь хүний хараа, сонсгол зэрэг хэд хэдэн мэдрэхүйг зэрэг хуурах боломжтой учраас илүү аюултай. Ялангуяа дүрс, дуу, уруулын хөдөлгөөн болон ярианы агуулга хоорондоо нийцсэн үед хэрэглэгч хуурамч эсэхийг ялгахад хүндрэлтэй болдог.



Зураг 1.9: Видео болон зургийн мэдээлэлд суурилсан хуурамч контент илрүүлэх ерөнхий схем

Зураг 1.9-д видео болон зургийн шинжүүдийг тус тусад нь боловсруулж, нэгтгэх давхаргаар нийлүүлэн эцсийн шийдвэр гаргах ерөнхий илрүүлэлтийн бүтцийг харуулав.

1.3 Аюулгүй байдлын эрсдэл, хор уршиг ба чиг хандлага

Хуурамч зураг, бичлэг, дууг хүмүүс үнэн гэж хүлээн авах, түүнд үндэслэн шийдвэр гаргах, итгэх, хуурагдах боломжтой. Иймээс энэ асуудал нь мэдээллийн аюулгүй байдал, хувь хүний эрх, байгууллагын нэр хүнд, цахим үйлчилгээний найдвартай байдал, биометрийн баталгаажуулалт, олон нийтийн итгэлцэл зэрэг олон талтай холбогдоно [8, 12].

Нийгмийн инженерчлэл ба залилан

Хуурамч контент нь хүний итгэл, сэтгэл зүйд нөлөөлөх замаар нийгмийн инженерчлэлийн халдлагыг илүү үнэмшилтэй болгодог. Өмнө нь залилан мэхлэлт ихэвчлэн зурвас, и-мэйл, утасны дуудлагад тулгуурладаг байсан бол одоо бодит хүнтэй төстэй зураг, бичлэг, дуу хоолой ашиглан хүний итгэх механизмыг чиглэсэн халдлага хийх боломж нэмэгдсэн [12].

2024 онд Хонконг дахь Arup компанийн ажилтан хуурамч видео уулзалтад оролцож, компанийн удирдлагууд мэт харагдсан хиймэл дүрүүдийн зааврааар мөнгөн шилжүүлэг хийсэн тохиолдол гарсан. Уг хэрэгт ойролцоогоор 25 сая ам.долларын хохирол учирсан гэж мэдээлсэн байна [16, 17].

Энэ жишээ нь нийгмийн инженерчлэлийн халдлага зөвхөн и-мэйл, утасны дуудлагаар хязгаарлагдахгүй, видео уулзалт болон хиймэл дүр төрх ашиглан байгууллагын итгэлцлийг хуурах түвшинд хүрснийг харуулж байна.

Зураг 1.10: Arup компанийн дипфейк видео хурлын залилангийн тухай мэдээний зураг

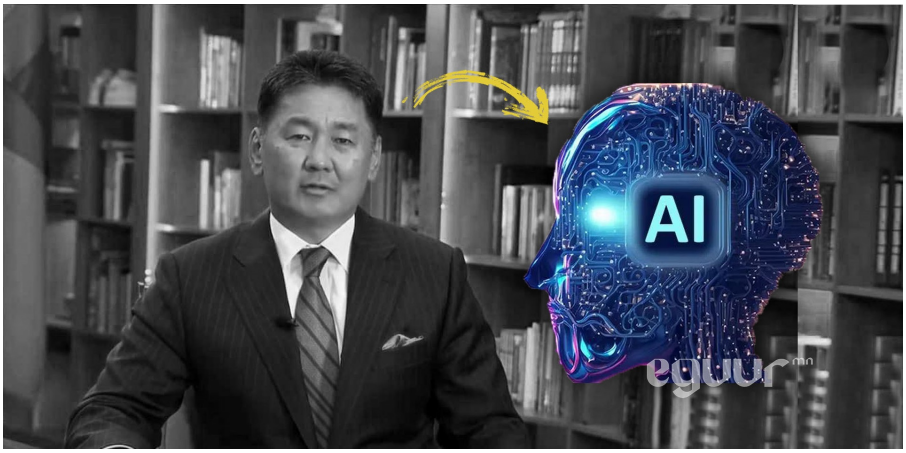
Зураг 1.10-д Arup компанитай холбоотой дипфейк залилангийн мэдээний зургийг харуулав.

Монголын дижитал орчин дахь бодит жишээ

Монгол Улсад интернэт болон нийгмийн сүлжээний хэрэглээ өргөн хүрээнд тархсан нь хиймэл зураг, видео, дуу хоолойд суурилсан мэдээлэл богино хугацаанд олон нийтэд түгэх нөхцөлийг бүрдүүлж байна. DataReportal-ийн 2026 оны тайланд Монгол Улсад 2.93 сая интернэт хэрэглэгч, 2.70 сая сошиал контент хэрэглэгчийн бүртгэл байсан гэж дурдсан нь цахим орчин дахь мэдээллийн нөлөөлөл өндөр байгааг харуулна [2].

Монголын цахим орчинд олны танил хүн, төрийн өндөр албан тушаалтан, байгууллагын удирдлагын дүр төрх, дуу хоолойг ашигласан хиймэл контент тархах нь олон нийтийн итгэл үнэмшил, байгууллагын нэр хүнд, мэдээллийн үнэн зөв

байдалд шууд нөлөөлдөг. Ийм төрлийн контентыг зөвхөн харагдах байдалд үндэслэн үнэлэх нь хангалтгүй бөгөөд эх сурвалж, нийтэлсэн орчин, өөрчлөлтийн шинж тэмдэг болон техникийн шалгалтыг хамтад нь авч үзэх шаардлагатай [18].



Зураг 1.11: Монгол Улсын Ерөнхийлөгч У.Хүрэлсүхийн дүр төрхийг ашигласан хиймэл контентын жишээ

Зураг 1.11-д төрийн өндөр албан тушаалтны дүр төрхийг ашигласан хиймэл контентын жишээг харуулав. Энэ нь Монголын цахим орчинд бодит болон хиймэл агуулгыг ялган таних, эх сурвалжийг нягтлах, автомат илрүүлэлт болон баримт шалгалтыг хослуулах шаардлагатайг илтгэнэ.

Хуурамч контентийн чиг хандлага

Хуурамч контентын хөгжил нь зөвхөн дүрс, дууны чанар сайжирч байгаагаар хязгаарлагдахгүй, харин түүнийг үүсгэх хэрэгсэл илүү хүртээмжтэй болж, залилан, дүр эсгэлт, худал мэдээлэл тараах хэлбэрүүдтэй нягт холбогдож байна. Deloitte-ийн тайланд дурдсанаар нийгмийн сүлжээн дэх дипфейк контентын тархалт 2019–2023 оны хооронд 550 хувиар өссөн гэж тэмдэглэсэн байна [19]. Иймээс хуурамч контентын асуудлыг зөвхөн технологийн шинэчлэл гэж бус, цахим итгэлцэл, мэдээллийн аюулгүй байдал, байгууллагын эрсдэлийн удирдлага, эрх зүй, ёс зүй болон олон нийтийн мэдээллийн боловсролтой холбоотой цогц асуудал гэж үзэх шаардлагатай. World Economic Forum-ийн 2025 оны тайланд худал болон төөрөгдүүлсэн мэдээлэл нь богино хугацааны дэлхийн гол эрсдэлүүдийн нэг болж байгааг онцолсон бөгөөд кибер орчны эрсдэлд дипфейк, хиймэл оюунаар дэмжигдсэн фишинг зэрэг халдлага багтаж байна.

1.4 Бүлгийн дүгнэлт

Энэ бүлэгт хуурамч контентын ойлголт, үүсгэх технологийн суурь болон аюулгүй байдлын эрсдэлийг нэгтгэн авч үзэв.

1. **Нэр томьёо ба ангилал.** Хуурамч контент, хиймэл контент, дипфейк гэсэн ойлголтуудыг ялгаж, зураг, видео, дуу болон олон модаль өгөгдөлд суурилсан хэлбэрүүдээр ангилан тайлбарлав.

2. **Технологийн үндэс.** GAN, авто-кодлогч, диффузийн загвар, трансформерт суурилсан архитектур зэрэг үүсгэгч аргууд нь хиймэл зураг, бичлэг, дуу хоолой үүсгэх үндсэн технологийн суурь болж байна. Эдгээр арга нь бодит мэт дүрслэл үүсгэхийн зэрэгцээ визуал, статистик болон акустик мөр үлдээдэг.

3. **Аюулгүй байдлын эрсдэл.** Дипфейк нь нийгмийн инженерчлэл, санхүүгийн залилан, нэр хүндэд халдах, биометрийн баталгаажуулалтыг хуурах, худал мэдээлэл тараах зэрэг эрсдэлтэй шууд холбогдож байна.

4. **Монголын нөхцөл.** Монгол хэлний дуу хоолой, дүрс бичлэгийн чанартай өгөгдөл хязгаарлагдмал байгаа нь хиймэл контент илрүүлэх судалгаанд хүндрэл үүсгэж байна. Энэ нь үндэсний хэл, орчны онцлогт тохирсон өгөгдөл, илрүүлэлтийн аргачлал боловсруулах хэрэгцээг харуулж байна.

Дүгнэж хэлбэл, хуурамч контентын асуудал нь зөвхөн технологийн хөгжил бус, мэдээллийн аюулгүй байдал, хувь хүний эрх, байгууллагын нэр хүнд болон үндэсний аюулгүй байдалтай шууд холбоотой. Иймээс зураг, видео, дуу зэрэг олон төрлийн мэдээллийг нэгтгэж авч үздэг олон модульт хандлага нь дипфейк илрүүлэлтнд чухал нөлөө үзүүлж байна.

БҮЛЭГ 2

Хуурамч контент илрүүлэлт ба баталгаажуулалтын арга, хэрэгсэл, алгоритмын харьцуулсан шинжилгээ

2.1 Хуурамч зураг, видео илрүүлэлтийн аргачлал ба нээлттэй эхийн хэрэгсэл, алгоритмын харьцуулалт

Хуурамч зураг, видео үүсгэх технологиуд нь боловсруулах зарчим, үлдээх хиймэл мөр, илрэх гажилтын хувьд хоорондоо ялгаатай. Тухайлбал, GAN, диффузийн загвар, авто-кодлогчид суурилсан нүүр солих, уруулын хөдөлгөөн тааруулах зэрэг аргууд нь дүрс болон видеог өөр өөр түвшинд өөрчилдөг. Иймээс хуурамч контент илрүүлэхдээ зөвхөн нэг шинж эсвэл нэг загварт тулгуурлах нь хангалтгүй бөгөөд хэд хэдэн төрлийн аргачлалыг хамтад нь авч үзэх шаардлагатай [6, 20].
Хуурамч зураг, видео илрүүлэх аргуудыг ерөнхийд нь дараах байдлаар ангилж болно. [5, 6, 20].

1. **Орон зайн шинжид суурилсан арга.** GAN, авто-кодлогч, диффузийн загвараар үүссэн зурагт пикселийн бүтэц, арьсны текстур, гэрэлтүүлэг, нүүрний ирмэг зэрэг харагдах түвшний зөрүү илэрдэг. Эдгээрийг CNN болон трансформерт суурилсан загвараар шалгана.

2. **Давтамжийн шинжид суурилсан арга.** Энэхүү арга нь контентийн давтамжийн тархалт, давтамжийн өөрчлөлт зэрэг нүдэнд шууд анзаарагдахгүй шинжийг ашигладаг. Ялангуяа GAN болон диффузийн загвараар үүссэн зурагт ийм ул мөр илэрдэг.

3. **Сэргээн босголтод суурилсан арга.** Дүрсийг дахин сэргээх үед бодит болон хиймэл зурагт ялгаатай алдаа гардаг. DIRE арга нь энэ зарчмаар диффузийн загвараар үүссэн зургийг илрүүлдэг.

4. **Хугацааны уялдаанд суурилсан арга.** Видео фреймүүдийн хооронд нүүрний хөдөлгөөн, уруулын хөдөлгөөний танилт, нүд анивчих байдал, гэрэлтүүлгийн тогтвортой байдлыг шалгана. Энэ нь нүүр солих, нүүрний хөдөлгөөн өөрчлөх аргаар үүссэн видеог илрүүлэхэд ашиглагдана.

5. **Нэгдсэн илрүүлэлтийн арга.** Орон зайн, давтамжийн, сэргээн босголтын болон хугацааны шинжийг хамтад нь ашиглана. Ингэснээр нэг шинжийн сул талыг бусад шинжээр нөхөж, илрүүлэлтийг илүү найдвартай болгодог.

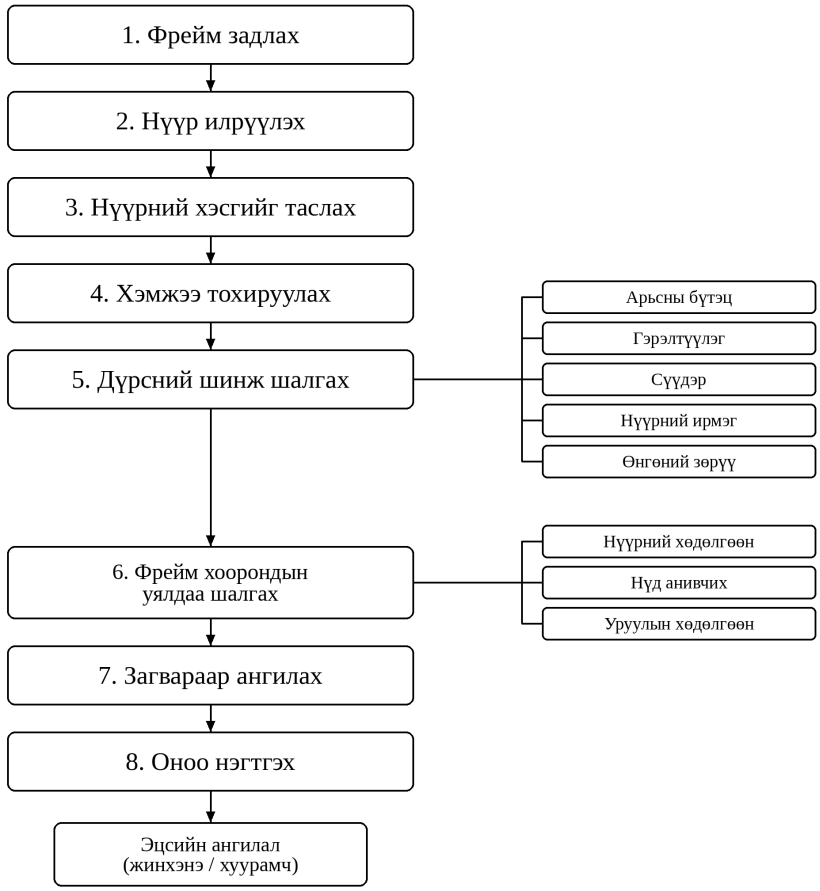
Зураг 2.1: Хуурамч зураг, видео илрүүлэх аргуудын ерөнхий ангилал

Зураг 2.1-д хуурамч зураг, видео илрүүлэхэд ашиглагддаг үндсэн аргачлалуудыг ерөнхийлөн харуулав.

Видео дипфейк илрүүлэгчийн аргачлал

Видео дипфейк илрүүлэлтэд видеог шууд бүхлээр нь ангилахаас өмнө фрейм задлах, нүүр илрүүлэх, нүүрний хэсгийг таслах, стандарт хэмжээтэй болгох зэрэг урьдчилсан боловсруулалт хийгддэг.

1. **Фрейм задлах.** Видео бичлэгээс тодорхой давтамжтайгаар фреймүүд ялган авна.
2. **Нүүр илрүүлэх.** Ялгасан фрейм бүрээс хүний нүүрний хэсгийг илрүүлнэ.
3. **Нүүрний хэсгийг таслах.** Илэрсэн нүүрийг фреймээс тусад нь авч, шаардлагагүй арын орчны нөлөөг багасгана.
4. **Хэмжээ тохируулах.** Нүүрний хэсгийг загварт оруулах стандарт хэмжээ, хэлбэрт оруулна.
5. **Дүрсний тогтвортой байдлыг шалгах.** Арьсны бүтэц, гэрэлтүүлэг, сүүдэр, нүүрний ирмэг, өнгөний зөрүү зэрэг шинжийг шинжилнэ.
6. **Фрейм хоорондын уялдаа шалгах.** Нүүрний хөдөлгөөн, нүд анивчих, уруулын хөдөлгөөн, толгойн байрлал зэрэг дараалсан фреймүүдэд тогтвортой байгаа эсэхийг үнэлнэ.
7. **Загвараар ангилах.** Бэлтгэсэн фреймүүдийг гүн сургалтын загварт оруулж, хуурамч байх магадлалыг тооцоолно.
8. **Оноо нэгтгэх.** Фрейм бүрийн үр дүнг нэгтгэн тухайн видеоны эцсийн оноо, ангиллыг гаргана.



Зураг 2.2: Видео дипфейк илрүүлэлтийн үндсэн үе шат

Зураг 2.2-д видео дипфейк илрүүлэлтийн үндсэн дарааллыг харуулав. Видео бичлэгээс фрейм ялган авч, нүүрний хэсгийг боловсруулсны дараа дүрсний шинж болон фрейм хоорондын уялдааг шалгаж, эцэст нь загварын оноогоор жинхэнэ эсвэл хуурамч гэж ангилна.

Хиймэл оюун ухаанаар үүсгэсэн зураг илрүүлэх аргачлал

Хиймэл оюун ухаанаар үүсгэсэн зургийг илрүүлэхдээ дүрсний бүтэц, гэрэлтүүлэг, сүүдэр, ирмэг, өнгөний зөрүүтэй байдал болон давтамжийн ул мөрийг шинжилдэг. Бодит зураг байгалийн гэрэл, камерын нөхцөлтэй уялдсан байдаг бол хиймэл зурагт хэт жигд бүтэц, хэлбэрийн алдаа, гэрэл сүүдрийн үл нийцэл илэрч болно. Хиймэл зураг илрүүлэх ерөнхий үе шат дараах байдалтай.

1. **Оролтын зураг.** Шинжлэх зургийг программд оруулна.
2. **Урьдчилсан боловсруулалт.** Зургийн хэмжээ, формат, өнгөний суваг болон пикселийн утгыг тохируулна.
3. **Дүрсний тогвортой байдлыг шалгах.** Арьсны бүтэц, ирмэг, гэрэлтүүлэг, сүүдэр, өнгөний зөрүү зэрэг харагдах шинжийг шалгана.
4. **Давтамжийн ул мөр шалгах.** Зурагт давтагдсан хиймэл хэв маяг, шахалтын үлдэгдэл, спектрийн өөрчлөлт байгаа эсэхийг шинжилнэ.
5. **Агуулгын нийцэл шалгах.** Нүүр, гар, нүд, шүд, үс, объектын хэлбэр, орчны байрлал зэрэг нь бодит зурагтай нийцэж байгаа эсэхийг шалгана.
6. **Оноо нэгтгэх.** Шалгасан шинжүүдийн үр дүнг нэгтгэн хиймэл байх магадлалыг тооцоолно.
7. **Эцсийн ангилал гаргах.** Нэгтгэсэн оноонд үндэслэн зургийг жинхэнэ, хуурамч эсвэл эргэлзээтэй гэж ангилна.

Зураг 2.3: Хиймэл оюун ухаанаар үүсгэсэн зураг илрүүлэх ерөнхий үе шат

Зураг 2.3-д хиймэл оюун ухаанаар үүсгэсэн зураг илрүүлэх ерөнхий үе шатыг харуулав.

Ашигласан нээлттэй эхийн хэрэгслүүд

Программ хангамжийн хэрэгжилтэд зураг, видео өгөгдөл боловсруулах, илрүүлэгч загвар ажиллуулах, үр дүнг үнэлэх, хадгалах болон тайлан гаргах нээлттэй эхийн хэрэгслүүдийг ашиглана.

1. **DeepFakeBench.** Видео дипфейк илрүүлэх checkpoint-үүдийг ашиглах, фрейм түвшний үнэлгээ хийх, клип түвшинд нэгтгэн эцсийн шийдвэр гаргах үндсэн benchmark орчин болгон ашиглана.

2. **PyTorch, torchvision, timm.** Xception, EfficientNet, F3Net, SPSL зэрэг CNN-д суурилсан загваруудыг ажиллуулах, checkpoint ачаалах, зураг болон видео фрейм дээр ангилалт хийхэд ашиглана.

3. **transformers, open-clip-torch.** Хиймэл оюун ухаанаар үүсгэсэн зургийн төлөөлөл, CLIP embedding болон GenImage төрлийн шинжийг боловсруулахад ашиглана.

4. **opencv-python, Pillow, imageio-ffmpeg.** Зураг унших, видеоноос фрейм ялгах, хэмжээ өөрчлөх, формат хөрвүүлэх, урьдчилсан боловсруулалт хийхэд ашиглана.

5. **facenet-pytorch, mediapipe.** Видео фреймээс нүүр илрүүлэх, нүүрний байрлал тогтоох, нүүрний хэсгийг тайрах болон landmark мэдээлэл боловсруулахад ашиглана.

6. **numpy, pandas, scipy, scikit-learn.** Илрүүлэлтийн оноо боловсруулах, feature нэгтгэх, accuracy, AUC зэрэг үнэлгээний үзүүлэлт тооцоход ашиглана.

7. **matplotlib, grad-cam.** Үр дүнгийн график гаргах, сэжигтэй бүсийг дүрслэх, загварын шийдвэрийг тайлбарлахад ашиглана.

8. **customtkinter.** Файл сонгох, зураг болон видео шинжлэх, илрүүлэлтийн оноо харуулах хэрэглэгчийн интерфэйс боловсруулахад ашиглана.

2.2 Өгөгдөл бэлтгэл ба үнэлгээний аргууд

Энэхүү судалгаанд хуурамч зураг, видео илрүүлэхэд ашиглах өгөгдлийг гурван үндсэн чиглэлээр авч үзсэн. Үүнд Хиймэл оюун ухаанаар үүсгэсэн зургийг илрүүлэх олон улсын өгөгдлийн сан, видео дипфейк илрүүлэх benchmark өгөгдөл болон Монгол хүний царайны датасет багтана.

Хүснэгт 2.1: Видео дипфейк илрүүлэхэд ашигласан өгөгдлийн сангууд

Өгөгдлийн сан	**Үүсгэсэн он**	**Өгөгдлийн хэмжээ**	**Ашигласан зорилго**
FaceForensics++	Rössler, 2019	5K видео, 500K+ фрейм	Үндсэн сургалтын өгөгдөл.
Celeb-DF v2	Li, 2020	6229 видео, 2M+ фрейм	Өндөр чанартай дипфейк дээр үнэлгээ.
DFDC	Facebook AI, 2020	100K+ клип	Том хэмжээний өгөгдөл дээр ерөнхий чадварыг үнэлэх.
DeeperForensics-1.0	Jiang, 2020	60K видео, 17.6M фрейм	Бодит орчны тогтвортой байдлыг үнэлэх.
Нийт: 171K+ видео, 20M+ фрейм			

Хүснэгт 2.1-д видео дипфейк илрүүлэхэд ашигласан гол өгөгдлийн сангуудыг харуулав. Эдгээр өгөгдөл нь DeepFakeBench-д хэрэгжсэн Xception, EfficientNet, F3Net, SPSL зэрэг checkpoint-үүдийг шалгах үндсэн benchmark орчин болсон.

Хүснэгт 2.2: Хиймэл оюун ухаанаар үүсгэсэн зураг илрүүлэхэд ашигласан өгөгдлийн сангууд

Өгөгдлийн сан	**Үүсгэсэн он**	**Өгөгдлийн хэмжээ**	**Ашигласан зорилго**
ForenSynths	Wang нар, CVPR 2020	Ойролцоогоор 724 мянган бодит болон хиймэл зураг	CNN болон GAN-аар үүсгэсэн зурагт үлдэх хиймэл оюун ухааны ул мөрийг шалгах.
GenImage	Zhu нар, 2023	2 саяас дээш бодит болон хиймэл зураг	Орчин үеийн хиймэл оюун ухаан зураг үүсгэгчдийн ялгааг харьцуулан шалгах.
DiffusionForensics / DIRE	Wang нар, 2023	Ойролцоогоор 464 мянган бодит болон диффузээр үүсгэсэн зураг	Диффуз загвараар үүссэн зургийг сэргээн босголтын ялгаагаар илрүүлэх.
Нийт ойролцоогоор: 3,188 мянгаас дээш зураг			

Хүснэгт 2.2-д Хиймэл оюун ухаанаар үүсгэсэн зургийг илрүүлэхэд ашиглах өгөгдлийн сангуудыг нэгтгэн харуулав. Эдгээр өгөгдөл нь CNNDetection, DIRE, UniversalFakeDetector

зэрэг зураг илрүүлэгч загваруудын ажиллах зарчмыг харьцуулахад ашигласан болно.

Зураг 2.4: Олон улсын дипфейк өгөгдлийн сангуудын жишээ

Зураг 2.4-д FaceForensics++, Celeb-DF v2 болон DeeperForensics-1.0 өгөгдлийн сангийн жишээг харуулав.
FaceForensics++ нь олон төрлийн нүүр хувиргалттай
Celeb-DF v2 нь өндөр чанартай дипфейк жишээтэй
DeeperForensics-1.0 нь гажуудал нэмсэн видео дипфейк илрүүлэхэд ашиглагддаг олон улсын benchmark өгөгдлүүд юм.

Өөрийн бэлтгэсэн өгөгдлийн сан

Олон улсын өгөгдлийн сангууд нь дипфейк илрүүлэлтийн судалгаанд өргөн хэрэглэгддэг боловч Монгол хүний царайны онцлог, зураг авах хэв маяг, хэвлэл мэдээлэлд ашиглагддаг нүүр зургийн хэлбэрийг бүрэн төлөөлөх боломж хязгаарлагдмал байдаг. Иймээс энэхүү программ хангамжийн хүрээнд Монгол хүний царайг агуулсан өөрийн туршилтын датасетийг бүрдүүлсэн.

1. Датасетийг Wikipedia эх сурвалжаас цуглуулсан бөгөөд *Mongolian politicians*, *Mongolian male judoka*, *Members of the State Great Khural* зэрэг ангиллуудыг ашигласан.

2. Нийт 459 хүний мэдээлэл цуглуулснаас MTCNN алгоритмаар 410 нүүрийг тогтвортой илрүүлсэн.

3. Илэрсэн нүүр бүрээс InceptionResnetV1 буюу VGGFace2 дээр урьдчилан сургасан загвараар 512 хэмжээст embedding гарган авсан.

4. Гарган авсан embedding-ийг царайны адилтгал хийх, мөн нэг хүний зураг сургалт, үнэлгээ, шалгалтын хэсэгт давхар орохоос сэргийлэхэд ашигласан.

5. 87 Монгол хүний бодит нүүрний зурагт тулгуурлан 540 өргөтгөсөн хувилбар болон 522 синтетик хуурамч зураг үүсгэсэн.

6. Эцэст нь нийт 1124 ширхэг 224×224 хэмжээтэй нүүрний crop зураг бүхий туршилтын датасет бэлтгэсэн.

Хүснэгт 2.3: Монгол царайны датасетийн бүтэц

Бүрэлдэхүүн	**Хэмжээ, хэлбэр**	**Ашиглах зорилго**
Монгол хүний мэдээлэл	Wikipedia болон Wikimedia Commons-оос цуглуулсан 459 хүний нэр, зураг	Монгол хүний царайны анхны сан бүрдүүлэх.
Илэрсэн нүүр	MTCNN алгоритмаар тогтвортой илэрсэн 410 нүүр	Царай илрүүлэх болон нүүрний crop үүсгэх үндсэн оролт болгох.
Царайны embedding	InceptionResnetV1 загвараар гаргасан 512 хэмжээст вектор	Царайны адилтгал хийх, Царай таних давхардлыг хянах.
Бодит нүүрний crop	87 хүний бодит нүүрнээс үүсгэсэн 540 augmentation зураг	Real ангиллын сургалт, баталгаажуулалт, тестэд ашиглах.
синтетик fake зураг	Дүрсний доройтуулах болон өөрчлөх боловсруулалтаар үүсгэсэн 522 зураг	Fake ангиллын жишээ болгон ашиглах.
Нийт: 1124 ширхэг 224×224 **хэмжээтэй нүүрний crop**		

Хүснэгт 2.3-д Монгол царайны датасетийн үндсэн бүтцийг харуулав. Уг датасет нь олон улсын өгөгдлийг орлох бус, харин Монгол хүний царайтай ойр туршилтын нэмэлт өгөгдөл болгон ашиглагдсан.

Хүснэгт 2.4: Монгол царайны датасетэд хадгалсан үндсэн мэдээлэл

Мэдээлэл	**Хадгалсан хэлбэр**	**Судалгаанд ашиглах зорилго**
Шошго	`label=real/fake`	Жинхэнэ болон хуурамч нүүрний хоёр ангиллын сургалт, тест хийх.
Дэд төрөл	`subtype`, `generator`	Хуурамч зураг ямар төрлийн боловсруулалтаар үүссэнийг тэмдэглэх.
Нүүрний crop	`224x224` хэмжээтэй зураг	Илрүүлэгч загварт нэг ижил хэмжээтэй оролт өгөх.
Царай таних мэдээлэл	`Царай таних`, 512 хэмжээст embedding	Нэг хүний зураг сургах, үнэлэх, шалгах хэсэгт давхар орохоос сэргийлэх.
Хуваалт	`split=сургах/val/шалгах`	Сургалт, баталгаажуулалт, тестийн үр дүнг тусад нь үнэлэх.
Эх сурвалж ба лиценз	`source`, `license`	Зургийн гарал үүсэл, ашиглах нөхцөлийг бүртгэх.

Хүснэгт 2.4-д Монгол царайны датасетэд хадгалсан үндсэн талбаруудыг харуулав. Эдгээр талбар нь жинхэнэ/хуурамч, нүүрний тасдалт, царай танил, сургах/үнэлэх/шалгах зэргийг харуулна.

Монгол царайны датасетийн дээж бүрийг manifest CSV файлд бүртгэсэн. Ингэхдээ `label`, `subtype`, `generator`, `identity`, `split`, `source`, `license` зэрэг талбарыг хадгалсан. Мөн сургах, үнэлэх, шалгах хуваалтыг хийхдээ нэг хүний зураг өөр өөр хэсэгт давхар орохгүй байх зарчмыг баримталсан.

Урьдчилсан боловсруулалт

Зураг болон видео өгөгдлийг илрүүлэгч загварт оруулахын өмнө нэг ижил бүтэцтэй болгож боловсруулсан. Видео өгөгдлөөс фрейм ялгаж, фрейм бүрээс нүүр илрүүлэн тасдаж авсан. Дараа нь нүүрний тасдалтуудыг тогтсон хэмжээ рүү хувиргасан.

Хиймэл зураг илрүүлэх өгөгдөлд зургийн хэмжээ өөрчлөх, формат жигдрүүлэх зэрэг боловсруулалт хийсэн. Видео дипфейк өгөгдөлд фрейм задлах, нүүр илрүүлэх, нүүрний хэсгийг тайрах, landmark болон mask мэдээлэл хадгалах алхмуудыг ашигласан. Харин Монгол царайны датасетэд MTCNN-ээр нүүр илрүүлж, 224×224 хэмжээтэй тасдалт үүсгэн, жинхэнэ болох хуурамч ангилалтай хадгалсан.

Үнэлгээний арга

Илрүүлэгч загварын үр дүнг зөвхөн нэг түвшинд бус, зураг, фрейм болон видео гэсэн гурван түвшинд үнэлсэн. Зургийн хувьд нэг дүрсийг шууд загварт оруулж бодит эсвэл хуурамч байх магадлалын оноог гаргасан. Харин видео өгөгдлийн хувьд бичлэгээс сонгосон фреймүүдийг тус бүрээр нь шинжилж, фрейм бүрийн илрүүлэлтийн оноог тооцсон. Дараа нь эдгээр фреймийн оноог нэгтгэн тухайн видеоны ерөнхий үнэлгээг гаргасан. Цаашилбал TP нь хуурамч контентыг жинхэнэ болон хуурамч гэж таньсан тохиолдол, TN нь бодит контентыг жинхэнэ гэж таньсан тохиолдол, FP нь бодит контентыг андуурч хуурамч гэж ангилсан тохиолдол, FN нь хуурамч контентыг андуурч бодит гэж ангилсан тохиолдлыг илэрхийлнэ. Загварын үр дүнг харьцуулахын тулд accuracy, precision, recall, F1-score болон AUC гэсэн үнэлгээний үзүүлэлтүүдийг ашигласан.

Хүснэгт 2.5: Үнэлгээний үзүүлэлтүүдийн тайлбар

Үзүүлэлт	**Томьёо / утга**	**Тайлбар**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Нийт өгөгдлийг жинхэнэ гэж ангилсан хувь.
Precision	$\frac{TP}{TP+FP}$	Хуурамч гэж ангилсан үр дүнгийн жинхэнэ хувь.
Recall	$\frac{TP}{TP+FN}$	Нийт хуурамч контентоос жинхэнэг илрүүлсэн хувь.
F1-score	$2 \times \frac{Precision \times Recall}{Precision+Recall}$	Precision болон recall-ийн тэнцвэржүүлсэн үзүүлэлт.
AUC	ROC муруйн доорх талбай	Бодит ба хуурамч контентыг ялгах ерөнхий чадвар.

Хүснэгт 2.5-д илрүүлэгч загварын гүйцэтгэлийг үнэлэхэд ашигласан үндсэн хэмжүүрүүдийг нэгтгэн харуулав. Accuracy нь нийт жинхэнэ ангилалтын түвшинг, precision нь хуурамч гэж тэмдэглэсэн үр дүнгийн найдвартай байдлыг, recall нь нийт хуурамч контентыг илрүүлэх чадварыг илэрхийлнэ. Харин F1-score нь precision болон recall-ийн тэнцвэрийг харуулдаг бол AUC нь бодит ба хуурамч контентыг ялгах ерөнхий чадварыг үнэлэхэд ашиглагдсан.

2.3 Контентын баталгаажуулалт ба дижитал мөр

Хуурамч контенттой тэмцэхэд **илрүүлэлт**, **баталгаажуулалт**, **гарал үүсэл** гэсэн гурван ойлголтыг ялгаж авч үзнэ. Илрүүлэлт нь зураг, видео, дууны доторх хиймэл мөрийг шинжилж бодит эсвэл хуурамч эсэхийг үнэлдэг. Баталгаажуулалт нь файл өөрчлөгдсөн эсэх, гарын үсэг, таних тэмдэг, хэш, метадата зэрэг илрүүлэлтээр шалгадаг. Харин гарал үүсэл нь тухайн контент хаанаас үүссэн, ямар хэрэгслээр боловсруулагдсан, ямар өөрчлөлт орсон талаар мэдээлэл өгөхөд чиглэнэ [21].

Дижитал мөр

Дижитал мөр нь контент дотор үлдэх техникийн ул мөр юм. Зурагт EXIF метадата, шахалтын түүх, пикселийн үлдэгдэл, давтамжийн хиймэл хэв шинж, GAN fingerprint, диффуз сэргээн босголтын зөрүү зэрэг шинжүүд хамаарна [5, 6].
Видео орчинд фрейм хоорондын зөрүү, нүүрний хөдөлгөөн, уруулын хөдөлгөөн, гэрэлтүүлгийн тогтвортой байдал, фрейм түвшний онооны хэлбэлзэл зэрэг шинжийг авч үздэг [20]. Дууны хувьд давтамжийн шинж, ярианы хэмнэл зэрэг нь шалгах боломжтой дохио болдог.

Эдгээр дижитал мөр нь дангаараа бүрэн нотолгоо болохгүй. Харин илрүүлэгч загварын оноо, метадата, гарал үүслийн мэдээлэлтэй хамт ашиглах үед тухайн контентын найдвартай байдлыг илүү үндэслэлтэй тайлбарлах боломж бүрдэнэ.

Гарал үүсэл ба баталгаажуулалт

Контентын баталгаажуулалтад харагддаг таних тэмдэг, үл харагдах таних тэмдэг, C2PA буюу Content Credentials, дижитал гарын үсэг, хэш шалгалт, метадата зэрэг аргууд ашиглагддаг [21].

Гэхдээ баталгаажуулалтын мэдээлэл байхгүй байна гэдэг нь тухайн контент заавал хуурамч гэсэн үг биш. Мөн баталгаажуулалтын мэдээлэл байгаа нь тухайн зураг, видеоны агуулга нь бодит гэсэн баталгаа болохгүй. Учир нь screenshot хийх, дахин хадгалах, сошиал сүлжээнд оруулах, форматыг өөрчлөх үед метадата болон таних тэмдэг алдагдаж болно. Иймээс баталгаажуулалт нь илрүүлэгчийг орлохгүй, харин нэмэлт баталгааны шалгуур болж ашиглагдана.

Hive AI-ийн API ашиглалт

Энэхүү судалгаанд нээлттэй эхийн загваруудаас гадна Hive AI-ийн гурван API-г нэмэлт шалгалтын хэрэгсэл болгон ашигласан. Үүнд Deepfake Detection API, Celebrity Recognition API, Likeness Detection API багтана [22].

Deepfake Detection API нь зураг болон видео дахь хүний нүүр бодит эсэх, эсвэл дипфейк шинжтэй эсэхийг итгэлцлийн оноогоор үнэлнэ. **Celebrity Recognition API** нь зурагт байгаа хүний нүүрийг өгөгдлийн сантай харьцуулж танихад ашиглагдана. **Likeness Detection API** нь зохиогчийн эрхээр баталгаажсан дүр болон, зохиолын дүр эсвэл брэндийн дүрслэлтэй төстэй байдал байгаа эсэхийг шалгахад хэрэглэгдэнэ.

Системд эдгээр API-ийн үр дүнг тус бүрийн итгэлцлийн оноо хэлбэрээр авч, бусад илрүүлэгч загварын үр дүнтэй харьцуулан ашигласан.

Хүснэгт 2.6: Hive AI API-уудын судалгаанд ашигласан үүрэг

API	**Шалгах зүйл**	**Судалгаанд ашигласан зорилго**
Deepfake Detection API	Зураг, видео дахь хүний нүүр бодит эсвэл дипфейк шинжтэй эсэх	Нүүрний бодит болон хуурамч байдлыг нэмэлт оноогоор шалгах.
Celebrity Recognition API	Зурагт байгаа хүний нүүр нийтийн танил хүнтэй тохирч байгаа эсэх	Хиймэл зурагт бодит хүний дүр төрх ашиглагдсан эсэхийг шалгах.
Likeness Detection API	Оюуны өмчөөр хамгаалагдсан дүр, баатар, брэндийн дүрслэлтэй төстэй байдал	Хиймэл оюун ухаанаар үүсгэсэн зурагт танигдахуйц дүрслэл ашиглагдсан эсэхийг шалгах.

Хүснэгт 2.6-д Hive AI-ийн гурван API-г судалгаанд ямар зорилгоор ашигласныг харуулав. Эдгээр API-ийн үр дүн нь нээлттэй эхийн загваруудын онооны хажуугаар нэмэлт шалгуур болж, дипфейк болон танил дүрийг ашигласан хиймэл контентыг илрүүлэхэд ашиглагдсан.

Хиймэл оюун ухаан платформын гарал үүслийн мэдээлэл

Хиймэл оюун ухаанаар үүсгэсэн контентыг баталгаажуулахад C2PA, Content Credentials, SynthID зэрэг арга хэрэглэгддэг. C2PA болон Content Credentials нь тухайн контент ямар хэрэгслээр үүссэн, засвар орсон эсэх, гарын үсгийн төлөв зэрэг метадата-г хадгалах зорилготой. SynthID нь контент дотор үл харагдах таних тэмдэг нэмж, тухайн агуулга Хиймэл оюун ухаанаар үүссэн эсэхийг шалгах боломж олгодог [21].

Хүснэгт 2.7: AI платформуудын гарал үүслийн мэдээллийн харьцуулалт

Платформ / стандарт	**Credential хэлбэр**	**Шалгаж болох мэдээлэл**
OpenAI / DALL·E 3	C2PA метадата, Content Credentials	AI-аар үүссэн эсэх, үүсгэсэн хэрэгсэл, C2PA manifest, гарын үсгийн төлөв.
Google DeepMind / SynthID	Үл харагдах таних тэмдэг	Зураг, видео, аудио, текст Хиймэл оюун ухаанаар үүссэн эсэхийг тусгай шалгагчаар шалгах.
Adobe Firefly / Content Credentials	C2PA-д суурилсан метадата	AI хэрэгсэл ашигласан эсэх, үүсгэсэн болон засварласан app, creator-ийн сонгосон нэмэлт мэдээлэл.
C2PA стандарт	Manifest, claim, signature	Үүсгэсэн болон засварласан үйлдэл, ашигласан эх материал, timestamp, үнэлэх status.

Хүснэгт 2.7-д Хиймэл оюун ухаан платформуудын гарал үүслийн мэдээллийг товч харуулав. Эдгээр мэдээлэл нь контентын гарал үүсэл болон засварын түүхийг шалгахад тусалдаг боловч тухайн агуулгын жинхэн эсэхийг дан ганц баталж чадахгүй. Иймээс илрүүлэлт, баталгаажуулалт, гарал үүслийн мэдээллийг нийтэд нь ашиглах илүү тохиромжтой.

2.4 Бүлгийн дүгнэлт

Энэхүү бүлэгт хийсэн ажлууд :

1. Хуурамч зураг, видео илрүүлэх аргуудыг орон зайн шинж, давтамжийн шинж, сэргээн босголтын алдаа, хугацааны уялдаа болон нэгдсэн илрүүлэлтийн зарчмаар ангилан тайлбарласан.

2. Хиймэл оюун ухаанаар үүсгэсэн зураг илрүүлэхэд CNNDetection, DIRE, UniversalFakeDetector болон CLIP-д суурилсан аргуудыг авч үзэж, тэдгээрийн ашиглах үндсэн шинжийг харьцуулан тайлбарлав.

3. Программд ашигласан DeepFakeBench, PyTorch, OpenCV, facenet-pytorch, scikit-learn, SQLite, customtkinter зэрэг нээлттэй эхийн хэрэгслүүдийн үүргийг тодорхойлов.

4. Хиймэл оюун ухаан зураг болон видео дипфейк илрүүлэх олон улсын өгөгдлийн сангуудыг нэгтгэн харьцуулж, ForenSynths, GenImage, DiffusionForensics, FaceForensics++, Celeb-DF v2, DFDC, DeeperForensics-1.0 өгөгдлийг судалгаанд ашиглахаар тусгав.

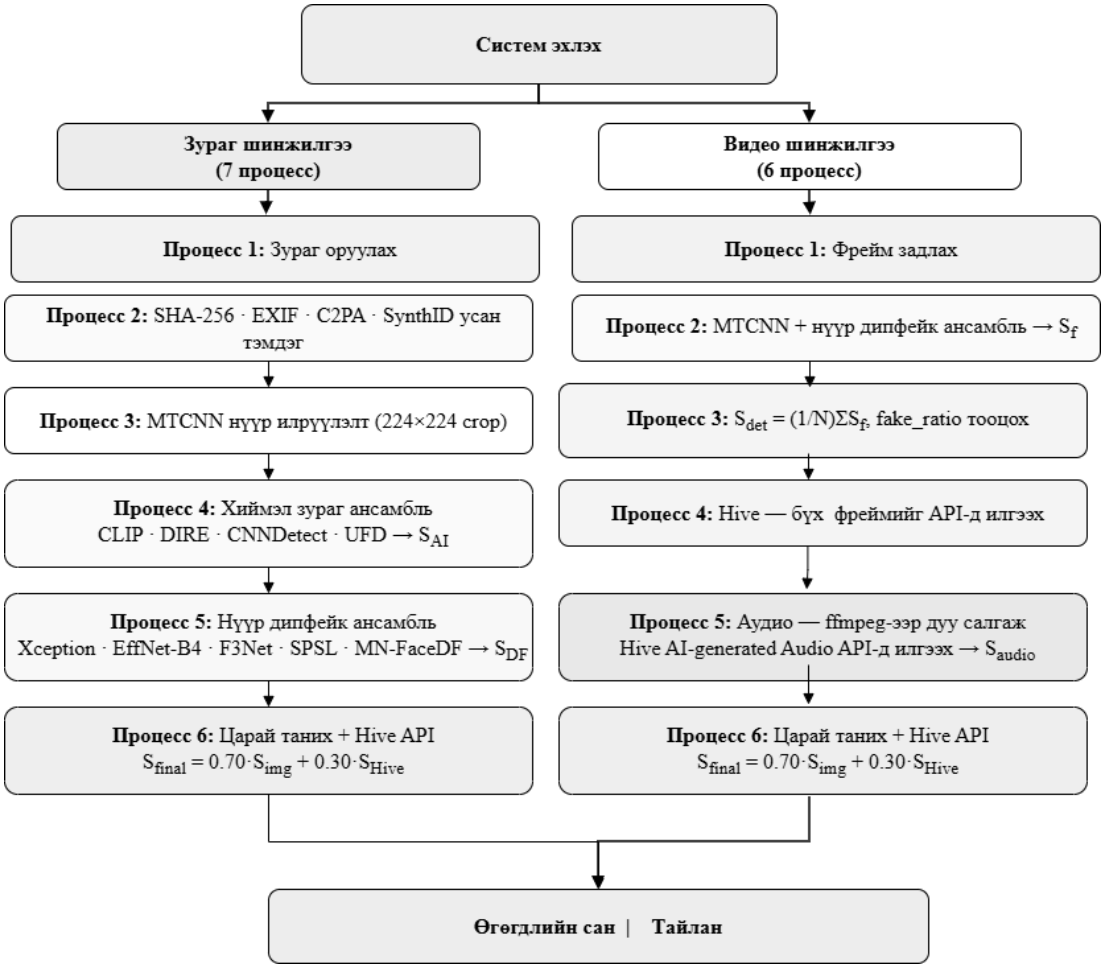
5. Өөрийн бэлтгэсэн өгөгдлийн санг бүрдүүлж, 459 хүний мэдээллээс 410 нүүр илрүүлэн, 1124 ширхэг 224×224 хэмжээтэй жинхэнэ болон хуурамч нүүрний тасдалт бэлтгэсэн.

6. Илрүүлэгчийн үр дүнг image-level, frame-level, video-level түвшинд үнэлэхээр тодорхойлж, accuracy, AUC, precision, recall, F1-score үзүүлэлтүүдийг ашиглахаар тусгав.

БҮЛЭГ 3

Техникийн хэрэгжилт ба интеграци

Дипломын ажлын хүрээнд хэрэгжүүлсэн программ нь хуурамч зураг, видеог илрүүлэхдээ олон загварыг хамтад нь ажиллуулдаг ансамбль зарчимтай. Системийн дотоод бүтэц нь зургаан үндсэн давхаргаас тогтоно: хэрэглэгчийн график интерфейс, систем эхлэх, зураг болон видеоны шинжилгээний хэсэг, загваруудын давхарга, гадаад API, хадгалалт болон тайлангийн давхарга. Зураг **??**-д тэдгээр давхаргын харилцан холбоосыг харуулав.



Зураг 3.1: Програмын архитектурын зураг

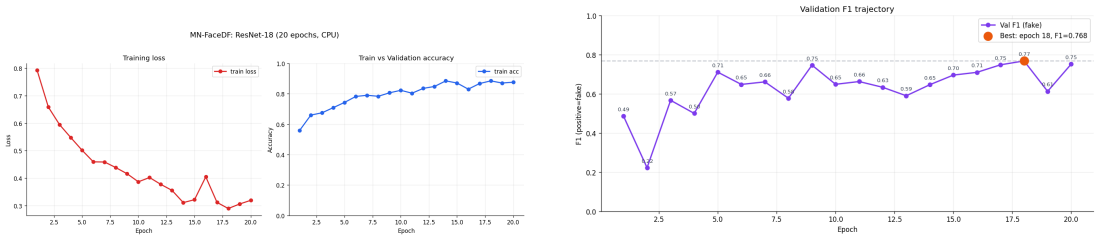
Зураг 3.1-д дипфейк илрүүлэх программын үндсэн давхаргууд болон тэдгээрийн хоорондын холбоог харуулав. Систем нь хэрэглэгчээс орж ирсэн зураг, видеог боловсруулж, илрүүлэгч загварууд болон Hive AI API-ийн үр дүнг нэгтгэн эцсийн үнэлгээ гаргана.

3.1 Илрүүлэгчийн загвар хэрэгжүүлэлт

MN-FaceDF загварын сургалт

Xception, EfficientNet, F3Net, SPSL зэрэг загварууд нь FaceForensics++ болон Celeb-DF дээр сургагдсан тул Монгол хүний нүүрний онцлогт тохирох туршилтын өгөгдөлгүй байсан. Программын нүүрэн дипфейк хэсгийг Монголын нөхцөлд тест хийхийн тулд MN-ImageDF датасет дээр **ResNet-18** загварыг дахин сургасан. Уг загварыг **MN-FaceDF** гэж нэрлэсэн.

Сургалтад ImageNet-1K дээр урьдчилан сургасан ResNet-18 жингийг ашиглан, эцсийн ангиллын давхаргыг жинхэнэ/хуурамч хоёр ангилалтай шинэ давхаргаар солисон. AdamW оптимайзер (lr=3×10^{-4}, weight_decay=10^{-4}), cross-entropy алдааны функц, 20 давталттай сургасан бөгөөд баталгаажуулалтын F1 оноо хамгийн өндөр байсан хувилбарыг checkpoint болгон хадгалсан.



(а) Алдааны муруй . (в) Баталгаажуулалтын F1 оноо давталт бүрд.

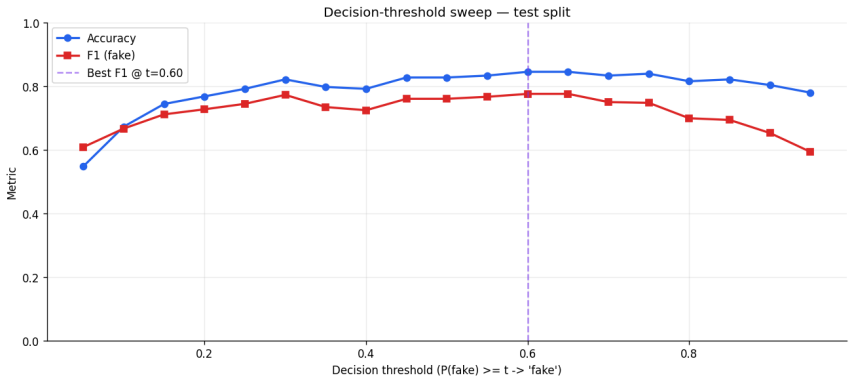
Зураг 3.2: MN-FaceDF загварын сургалтын явц .

Зураг 3.2-д MN-FaceDF загварын сургалтын алдааны муруй болон F1 үзүүлэлтийн муруйг хамт харуулав. Сургалт ахих тусам алдаа буурч, F1 үзүүлэлт тогтворжиж байгаа нь загвар жинхэнэ/хуурамч ангиллын онцлогийг сурч чадсаныг илтгэнэ.

MN-FaceDF загварын тестийн үр дүн

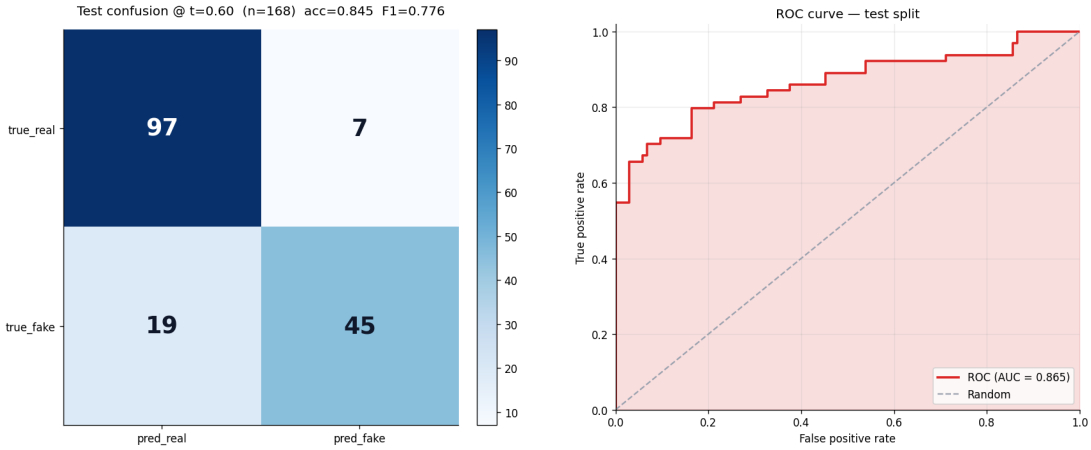
Зураг 3.3: тестийн жинхэнэ ба хуурамч ангиллын онооны тархалт (n=168).

Зураг 3.3-д тестийн 168 дээжийн хуурамч байх магадлалын тархалтыг жинхэнэ (Real) ба хуурамч (Fake) ангиллаар тусгаж харуулав. Хоёр ангиллын тархалт бие биенээсээ тодорхой зайтай давхцах нь $\tau = 0.60$ босгоор зохистой ангилах боломжтойг харуулна.



Зураг 3.4: MN-FaceDF: босго утгын өөрчлөлтөөс хамаарсан precision, recall, F1, accuracy.

Зураг 3.4-д босго утга τ-г 0-оос 1 хүртэл өөрчлөхөд precision, recall, F1 болон accuracy хэрхэн хувирч буйг харуулав. Precision нэмэгдэхэд recall буурч, эсрэгээрээ ажиллах нь стандарт хэв маяг. $\tau = 0.60$ утгад F1 болон accuracy хоёулаа тэнцвэртэй байдаг тул энэ утгыг программд ашиглахаар сонгосон.

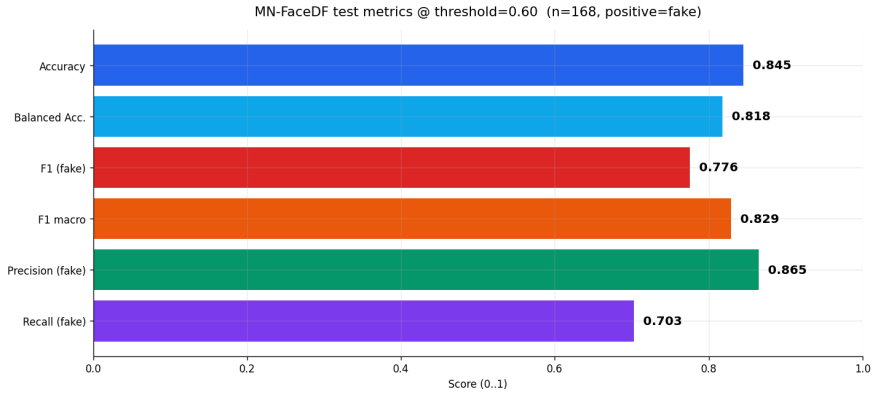


(а) Алдааны матриц, $\tau = 0.60$.

(в) ROC муруй.

Зураг 3.5: MN-FaceDF загварын алдааны матриц ($\tau = 0.60$, зүүн) ба ROC муруй (баруун).

Зураг 3.5-д $\tau = 0.60$ босгоор тооцсон алдааны матриц болон ROC муруйг хамт харуулав. Алдааны матрицын зэрэгчилсэн утгууд (жинхэнэ гэж зөв таних, хуурамч гэж зөв таних) нь бусдаасаа давамгайлж байгаа нь загвар тогтвортой ажиллаж байгааг харуулна. ROC AUC 0.865 нь хоёр ангиллыг ялгах ерөнхий чадварыг илэрхийлдэг.



Зураг 3.6: MN-FaceDF загварын тестийн үзүүлэлтүүдийн нэгтгэл ($\tau = 0.60$, n=168).

Зураг 3.6-д $\tau = 0.60$ босгоор тооцсон Accuracy, Balanced Accuracy, Precision, Recall, F1 болон ROC AUC үзүүлэлтүүдийг баганан диаграмаар нэгтгэн харуулав. Тооцоолсон үзүүлэлтүүдийг Хүснэгт 3.1-д тоон байдлаар оруулав.

Хүснэгт 3.1: MN-FaceDF загварын тестийн үнэлгээ ($\tau = 0.60$, 168 дээж)

Хэмжүүр	**Утга**	**Тайлбар**
Accuracy	0.845	168 тест дээжээс зөв ангилсан хувь.
Balanced Accuracy	0.818	жинхэнэ ба хуурамч утгыг тэнцвэртэй авч үзсэн утга.
Precision (fake)	0.865	Хуурамч гэснийх нь үнэхээр хуурамч байх хувь.
Recall (fake)	0.703	Бодит хуурамчуудын хэдэн хувийг олж илрүүлсэн.
F1 (fake)	0.776	Precision ба recall-ийн тэнцвэржсэн дундаж.
Macro F1	0.829	Ангилал тус бүрд тооцоод дундажилсан.
ROC AUC	0.865	Босго өөрчлөгдсөн нөхцөлд ялгах чадвар.

Хүснэгт 3.1-д MN-FaceDF загварын 168 тест дээжийн үнэлгээний үзүүлэлтийг $\tau = 0.60$ босгоор нэгтгэн харуулав. Accuracy 0.845, хуурамч ангиллын F1 0.776, ROC AUC 0.865 үзүүлсэн нь Монгол царайны датасет дээр сургасан загвар жинхэнэ болон хуурамч нүүрийг тодорхой хэмжээгээр ялгах боломжтойг харуулж байна. Тестийн алдааны матриц: жинхэнэ зөв таних 93, хуурамч зөв таних 46, хуурамч буруу таних (FN) 18, жинхэнэ буруу таних (FP) 11.

Гадаад датасет дээрх загваруудын харьцуулалт

Программд интеграц хийсэн Xception, EfficientNet-B4, F3Net, SPSL загварууд нь DeepFakeBench-ийн урьдчилан сургасан checkpoint хэлбэрээр ашиглагдсан.

Хүснэгт 3.2: Гадаад датасет дээрх илрүүлэгч загваруудын AUC харьцуулалт

Датасет	**Xception**	**EffNet-B4**	**F3Net**	**SPSL**
FaceForensics++ (c23)	0.9637	0.9567	0.9635	0.9610
Celeb-DF v2	0.7365	0.7487	0.7352	0.7650
DFDC	0.7077	0.6955	0.7021	0.7040
DeeperForensics-1.0	0.8341	0.8330	0.8431	0.8767
UADFV	0.9379	0.9472	0.9347	0.9424

Хүснэгт 3.2-д Xception, EfficientNet-B4, F3Net, SPSL загваруудын олон улсын датасет дээрх AUC үзүүлэлтийг нэгтгэн харуулав. Сургалтын үндсэн өгөгдөл болсон FaceForensics++ дээр бүх загвар 0.96 орчим AUC үзүүлэх боловч, cross-dataset тест (Celeb-DF, DFDC) дээр үзүүлэлт 0.70–0.75 хүртэл буурдаг. Иймд нэг загварт найдалгүйгээр олон илрүүлэгчийг хослуулах шаардлагатай гэдгийг харуулсан.

Аудио дипфейк илрүүлэлт (Hive AI)

Программын аудио дипфейк илрүүлэлтийг **Hive AI-generated Audio API** ашиглан гүйцэтгэнэ. Видео файлаас салгасан дуу авиа эсвэл дан аудио файлыг Hive API-д илгээж, хүний дуу хоолой жинхэнэ эсвэл TTS/дуу хоолой хуулбарлах аргаар үүсгэгдсэн эсэхийг итгэлцлийн оноогоор үнэлдэг. Энэ ажиллагааг видео шинжилгээний Шат 5-д дэлгэрэнгүй харуулав.

3.2 Хуурамч зураг, бичлэг илрүүлэх програмын хөгжүүлэлт

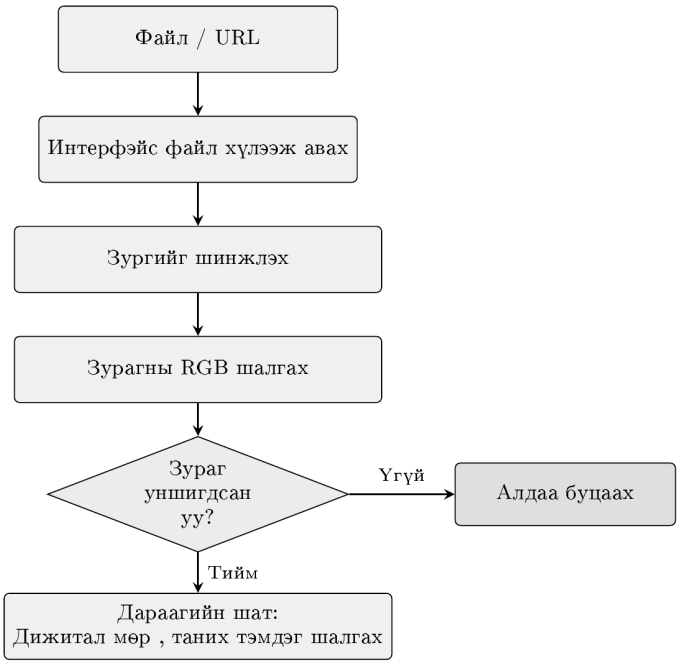
Зураг шинжилгээ

Зураг шинжлэх функц нь оролтоос эцсийн дүгнэлт хүртэл долоон үе шатаар дамждаг. Зураг 3.7-д шат бүрийг гарчигтай зурвас болгож, доторх бүрэлдэхүүн хэсэг бүрийг тусдаа дөрвөлжинд харуулав.

Зураг 3.7: Зураг шинжилгээний боловсруулалтын 7 шат, шат бүрийн бүрэлдэхүүн хэсэг тус бүрээр

Зураг 3.7-д зураг шинжилгээний 7 шатыг харуулав. Шат бүрийн зүүн талд гарчиг, баруун талд тухайн шатны бүрэлдэхүүн хэсгүүдийг дөрвөлжинд оруулсан. Дараах хэсгүүдэд шат бүрийг дэлгэрэнгүй тайлбарлав. Хэрэгжилтийн кодыг Хавсралт А-д жагсаав.

Шат 1: Оролтын зураг. Хэрэглэгч локал, чирж тавих эсвэл URL-аар зураг оруулна. Программ зурагны RGB-г унших ба амжилтгүй бол алдаа гаргаж зогсоно.



Зураг 3.8: Шат 1 — Зургийг ачаалах.

Зураг 3.8-д зураг хүлээн авах, зурагны RGB унших ба гэмтсэн файлыг таслан зогсоох шийдвэрийн дарааллыг блок схемээр харуулав.

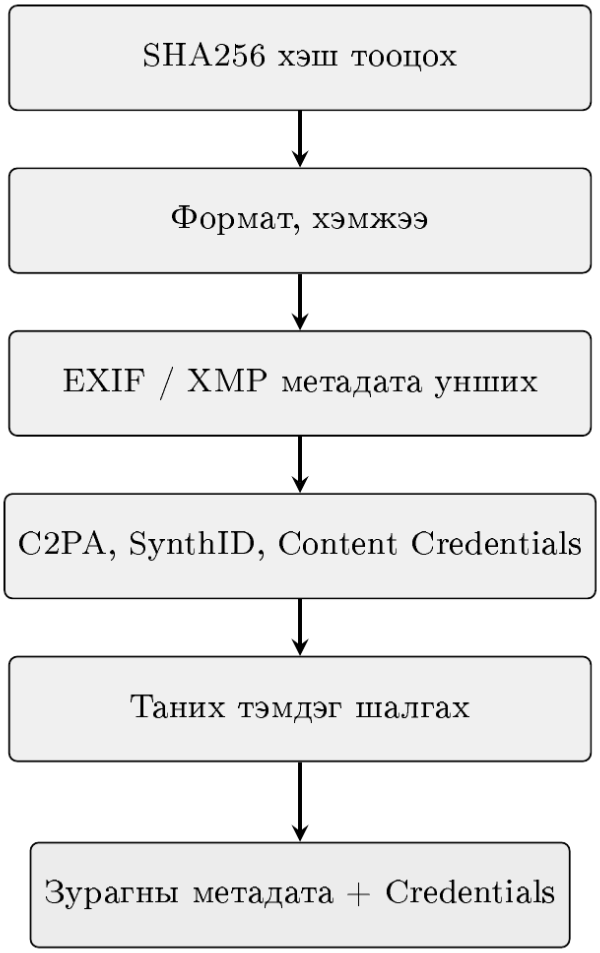
Шат 2: Дижитал мөр, таних тэмдэг шалгах зурагт EXIF, XMP, C2PA, SynthID, таних тэмдэг зэрэг гарал үүслийн тэмдэг байгаа эсэхийг шалгана. Илэрсэн тэмдгийг нэмэлт нотолгоо болгон Илрүүлэгчийн үр дүн дээр хадгална.

Зураг 3.9-д SHA256 хэш-аас эхлээд формат, EXIF/XMP, C2PA, SynthID, таних тэмдэг хүртэлх гарал үүслийн мөрийг шинжлэн процессыг блок схемээр харуулав.

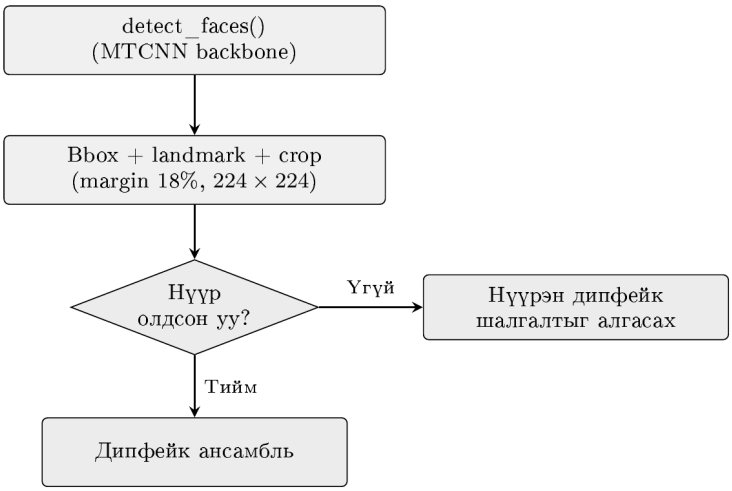
Шат 3: Царай илрүүлэлт.
FaceExtractor нь **MTCNN**-ийг ашиглан нүүрний bbox-ийг олж, margin нэмж 224× 224 хэмжээтэй тасдаж бэлэн болгоно. Нүүр илрээгүй бол нүүрэн дипфейк шалгалтыг алгасаж, зөвхөн бүхэл зургийн илрүүлэгчид ажиллана.

Зураг 3.10-д MTCNN-ээр нүүрийг олж, margin нэмж 224×224 crop болгох ба нүүр олдоогүй үед дипфейк шалгалтыг алгасах салаалалтыг харуулав.

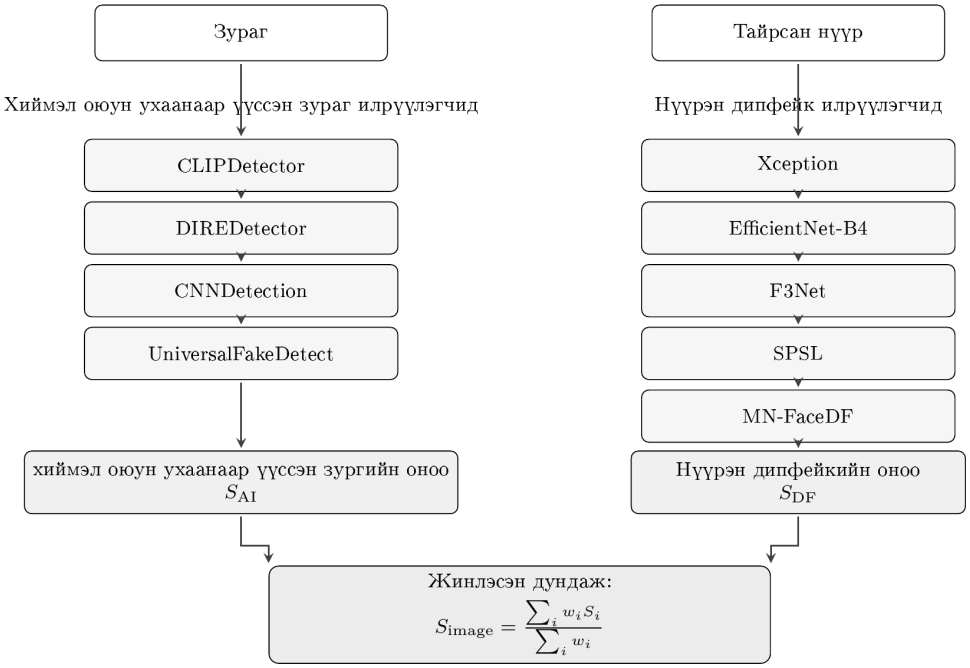
Шат 4–5: Хиймэл оюун ухаанаар үүссэн зураг ба нүүрэн дипфейк ансамбль. Зургийн хувьд CLIPDetector, DIREDetector, CNNDetection, UniversalFakeDetect загварууд хиймэл оюун ухаанаар үүсгэгдсэн зурагт үлдсэн орон зайн, давтамжийн болон сэргээн босголтын шинжийг шалгана. Хэрэв нүүр илэрсэн бол Xception, EfficientNet-B4, F3Net, SPSL, MN-FaceDF загварууд тайрсан нүүрэн дээр нэмэлт шалгалт хийнэ. Бүх илрүүлэгчийн оноог жинлэсэн дундажаар нэгтгэж S_{image} оноо гаргана.



Зураг 3.9: Шат 2 — Дижитал мөр, таних тэмдэг шалгах.



Зураг 3.10: Шат 3 — Царай илрүүлэлт.



Зураг 3.11: Шат 4–5 — хиймэл оюун ухаанаар үүссэн зураг илрүүлэгчид ба нүүрэн дипфейк ансамбль.

Зураг 3.11-д бүхэл зургийн хиймэл оюун ухаанаар үүссэн илрүүлэгчид (зүүн) ба тайрсан нүүрэн дээрх дипфейк илрүүлэгчид (баруун) зэрэгцэн ажиллаж, жинлэсэн дундажаар S_{image} оноо болон нэгтгэгдэх бүтцийг харуулав.

Шат 6: Царайг таних Илэрсэн нүүрнүүдийг лавлах мэдээлэлтэй харьцуулж, төстэй царай байгаа эсэхийг хайна. Илэрсэн царайг нэмэлт мэдээлэл болгон бүртгэнэ. Hive API key тохируулсан тохиолдолд Deepfake / Celebrity / Likeness API-уудыг дуудаж S_{HiveAPI} оноог буцаана.

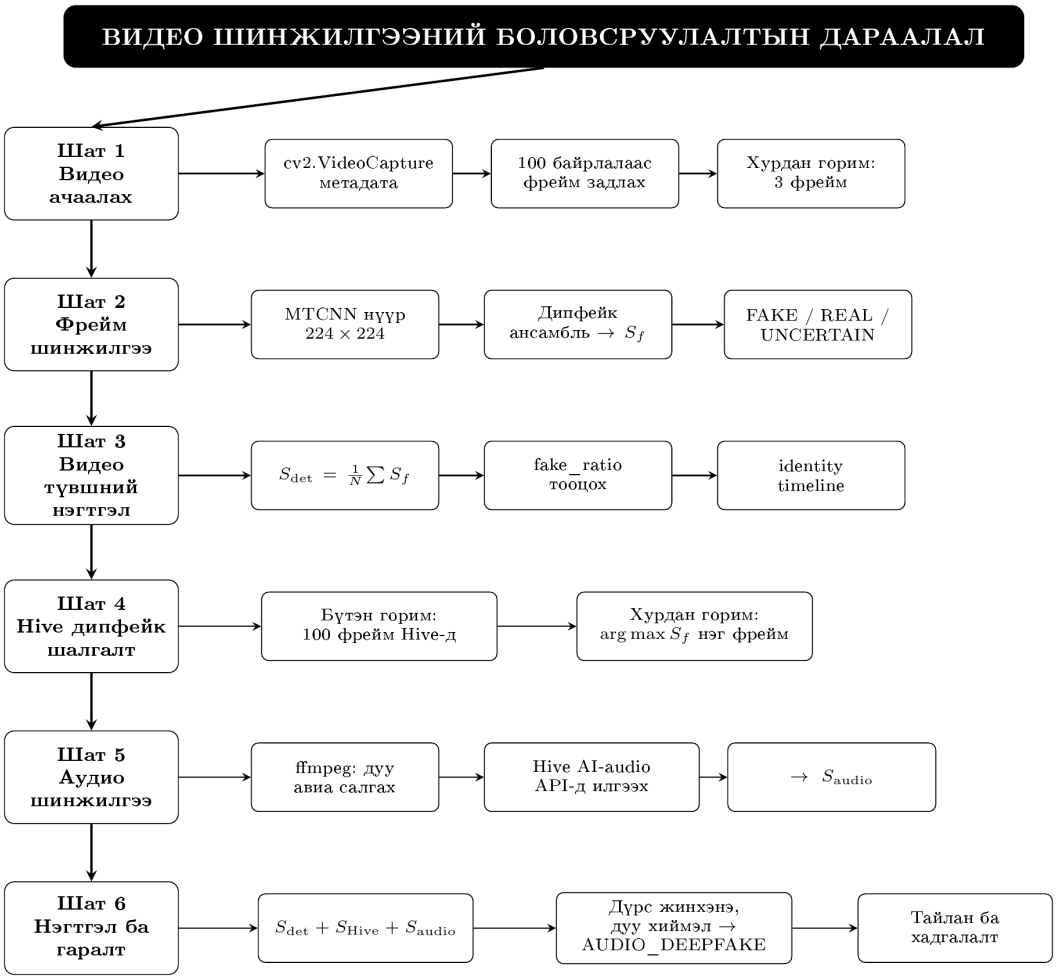
Шат 7–8: Эцсийн оноо ба үр дүн. ансамбльийн S_{image} оноог Hive API-ийн S_{HiveAPI}-тэй 0,70 : 0,30 жинтэйгээр нэгтгэж S_{final} оноо тооцоод дүгнэлт гаргана.

Хүснэгт 3.3: Зураг шинжилгээний дүгнэлт гаргах босго утгууд

Нөхцөл	**Дүгнэлт**
$S_{\text{final}} \geq 0.80$	Хуурамч байх өндөр магадлалтай.
$0.60 \leq S_{\text{final}} < 0.80$	Хуурамч эсвэл хиймэл оюун ухаанаар үүсгэгдсэн магадлалтай.
$0.40 \leq S_{\text{final}} < 0.60$	Тодорхойгүй, нэмэлт шалгалт шаардлагатай.
$S_{\text{final}} < 0.40$	Бодит байх магадлалтай.

Видео шинжилгээ

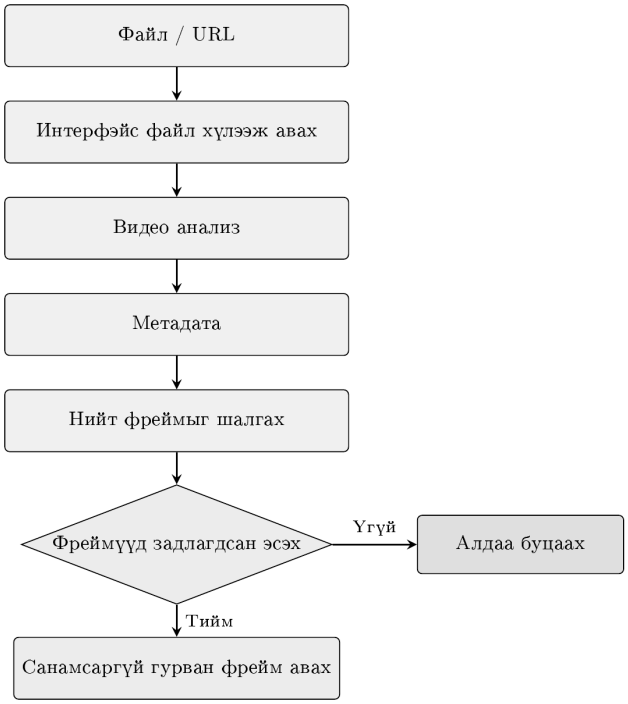
Видео шинжилгээ нь нийт фрейм бүрийг боловсруулж зургаан үндсэн шатаар шалгадаг. Зураг 3.12-д шат бүрийн бүрэлдэхүүн хэсгийг дөрвөлжинд харуулав.



Зураг 3.12: Видео шинжилгээний боловсруулалтын 6 шат, шат бүрийн бүрэлдэхүүн хэсэг тус бүрээр

Зураг 3.12-д видео шинжилгээний 6 шатыг харуулав. Видео ачаалах, фрейм шинжилгээ, видео түвшний нэгтгэл, Hive шалгалт, аудио шинжилгээ болон эцсийн нэгтгэлт гэсэн шатуудаас бүрдэнэ.

Шат 1: Оролтын видео. cv2.VideoCapture-ээс нийт фрейм, хугацаа уншаад, нийт фреймээс давхцахгүй санамсаргүй байрлал дахь гурван фреймийг сонгож, тус бүрд фреймыг шинжилнэ.

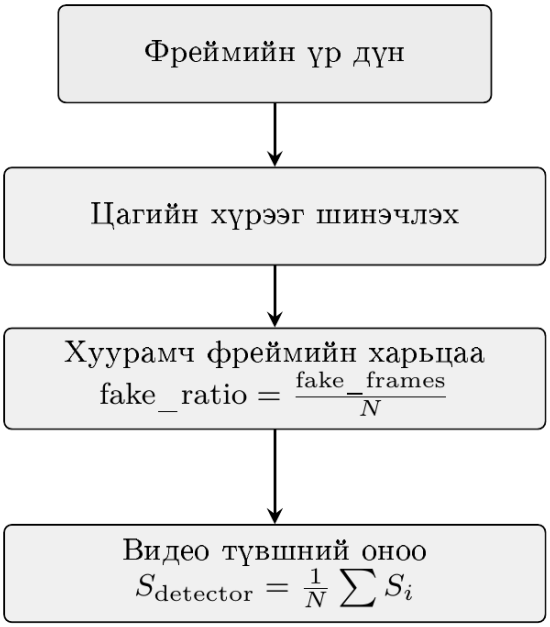


Зураг 3.13: Шат 1 — Оролтын видео.

Зураг 3.13-д видео хүлээн авч метадата уншаад санамсаргүй байрлалын гурван фреймийг ялгах ба гэмтсэн файлыг таслах дарааллыг харуулав.

Шат 2: Фрейм бүрийн шинжилгээ. Сонгогдсон фрейм бүр дээр нүүрийг илрүүлж тайрна. Тайрсан нүүрийг Xception, EfficientNet-B4, F3Net, SPSL, MN-FaceDF загваруудад дамжуулж, оноог нь дундажилж фреймийн оноо S_f-ийг гаргана. $S_f \geq 0.60$ бол FAKE, $S_f < 0.51$ бол REAL, хооронд нь UNCERTAIN гэж тэмдэглэнэ.

Шат 3: Видео түвшний нэгтгэл. Фрейм бүрийн оноог дундажилж видеоны S_{detector} оноог, хуурамч фреймийн харьцааг (fake_ratio) болон identity timeline-ийг гаргана.



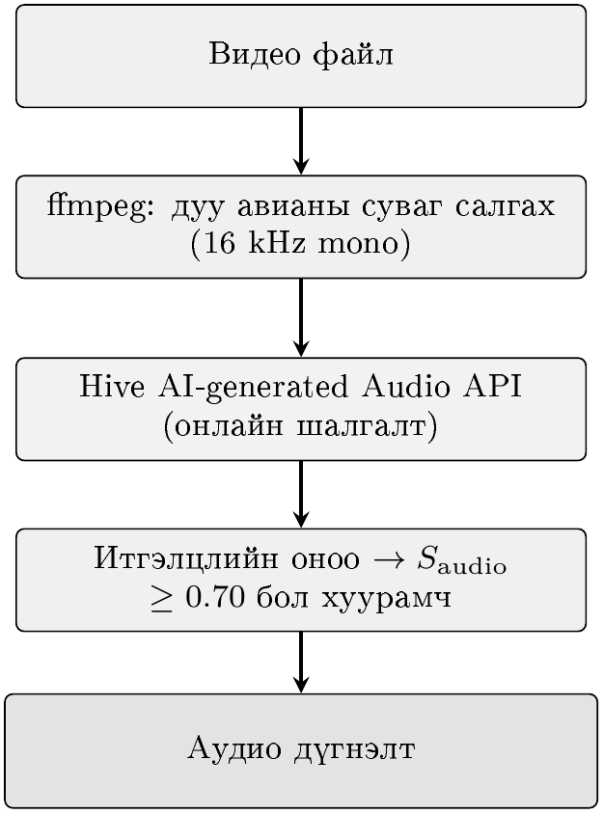
Зураг 3.14: Шат 3 — видео түвшинд оноо нэгтгэх.

Зураг 3.14-д фрейм бүрийн оноог дундажилж видео түвшний S_{detector} оноо, хуурамч фреймийн харьцаа болон фреймийн хүрээг шинэчлэх дарааллыг харуулав.

Шат 4: Hive хиймэл оюун ухаанаар үүссэн шалгалт. Hive API нь нэг видеон дээр зөвхөн хамгийн сэжигтэй фреймийг ($\max S_f$) ашиглан хиймэл оюун ухаан/дипфейк шалгалт хийнэ. API key тохируулаагүй бол алгасагдана.

Шат 5: Аудио дипфейк шинжилгээ .

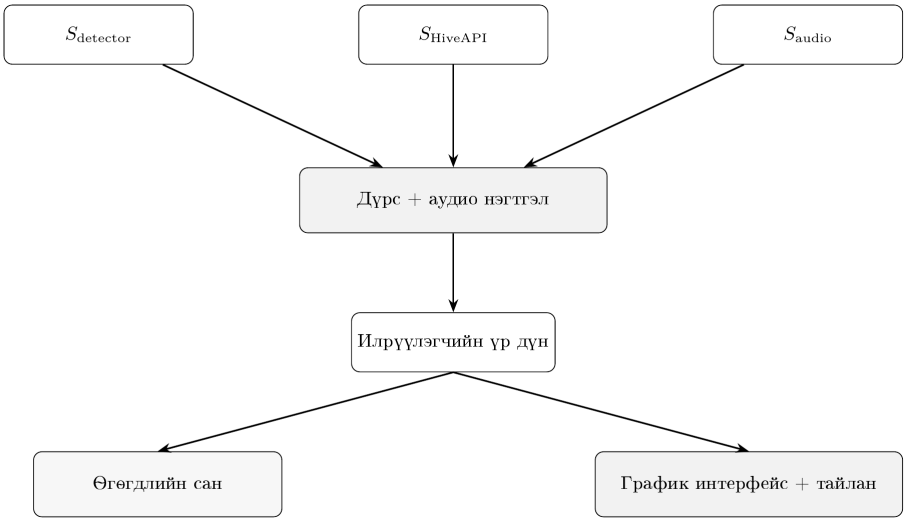
Видеоны дуу авианы сувгийг ffmpeg-ээр салгаж 16 kHz mono болгоод, **Hive AI-generated Audio API**-д илгээж үнэлнэ. Hive нь хүний дуу хоолой жинхэнэ эсвэл TTS, дуу хоолой хуулбарлах аргаар үүсгэгдсэн эсэхийг итгэлцлийн оноогоор буцаадаг бөгөөд ≥ 0.70 бол хуурамч гэж тэмдэглэн эх үүсвэрийн engine-ийг тогтоодог. Дуу авиагүй видеог зөв таниж алгасна.



Зураг 3.15: Шат 5 — видео доторх аудио дипфейк шинжилгээ (Hive AI).

Зураг 3.15-д видеоны дуу авианы сувгийг ffmpeg-ээр салгаж, Hive AI-audio API-аар үнэлж аудио дүгнэлт гаргах дарааллыг харуулав. Дуу авиагүй видеог зөв таньж тусгай `AudioTrackVerdict(has_audio=False)` төлөв буцаадаг.

Шат 6: Дүрс + аудио нэгтгэл ба гаралт. Дүрсний S_{detector}, Hive фреймийн S_{HiveAPI} ба аудио S_{audio} оноог нэгтгэж эцсийн дүгнэлт гаргана. **Дүрс жинхэнэ ч дуу нь хиймэл** бол видеог AUDIO_DEEPFAKE болгон тэмдэглэдэг нь чухал онцлог. Үр дүнг өгөгдлийн санд хадгалж, график интерфейст харуулах болон тайлан болгон гаргана.

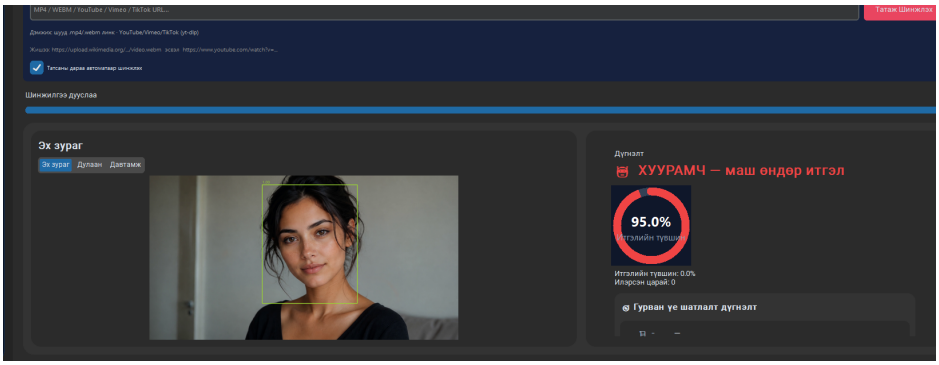


Зураг 3.16: Шат 6 — дүрс ба аудиог нэгтгэн эцсийн дүгнэлт, хадгалалт.

Зураг 3.16-д дүрсний, Hive фреймийн болон аудионы оноог нэгтгэн эцсийн дүгнэлт гаргаж, үр дүнг өгөгдлийн санд хадгалан график интерфейст харуулах болон тайлан болгон гаргах схемийг харуулав.

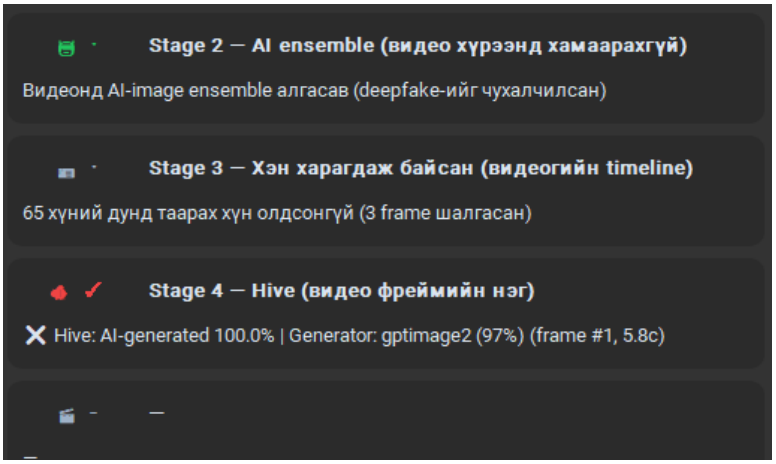
3.3 Хэрэгжүүлэлтийн үр дүн

Программ ажиллаж эхлэсний дараа хэрэглэгч зураг, видео оруулж шинжилгээ эхлүүлэх боломжтой график интерфэйс харагдана.



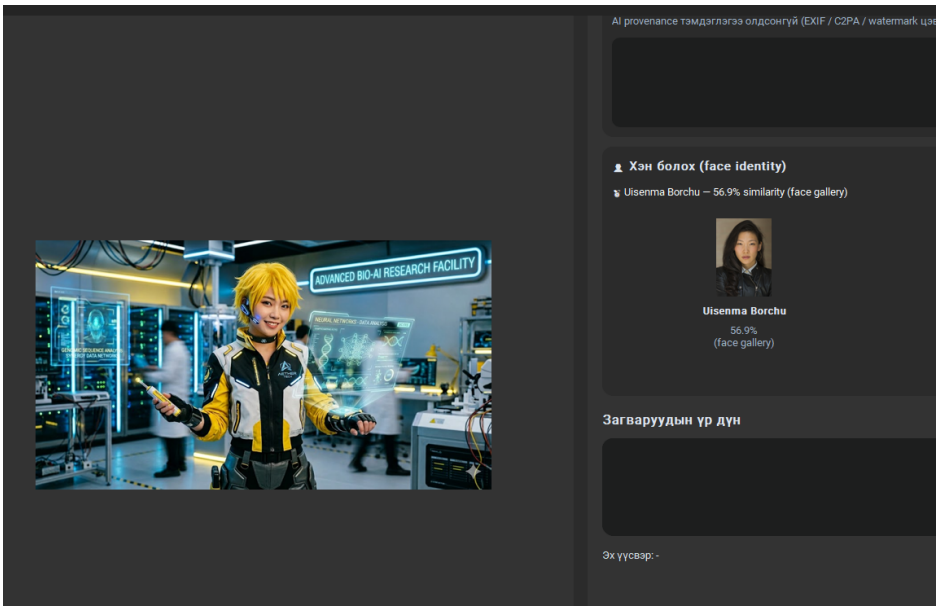
Зураг 3.17: Видео шинжилгээний оролтын интерфейс.

Зураг 3.17-д хэрэглэгч видео файл оруулах болон URL-аас татах оролтын интерфейс, шинжилгээний явцыг харуулав.



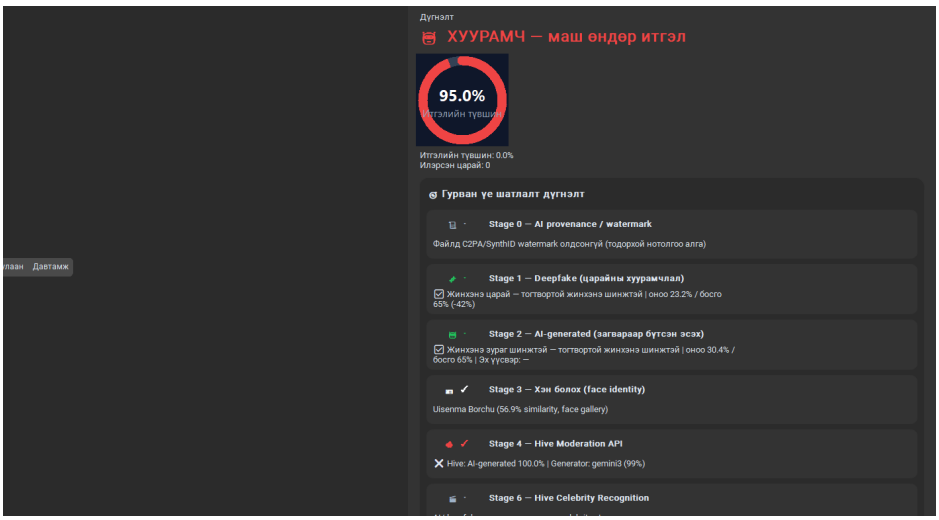
Зураг 3.18: Видео шинжилгээний үе шаттай дүгнэлт.

Зураг 3.18-д видео шинжилгээний үе шаттай дүгнэлт болон хиймэл оюун ухаанаар үүссэн ансамбль, царай таних, Hive шалгалтын дэлгэрэнгүй задаргааг харуулав. Зураг 3.18-д хиймэл оюун ухаанаар үүссэн ансамбль болон царай таних шалгалтын үр дүн, Зураг 3.15-д тухайн видеоны дуу авианы Шат 5 дүгнэлт (жинхэнэ/хиймэл дуу, эх үүсвэр, сэжигтэй сегмент, Hive AI-audio хувь) харагдана.



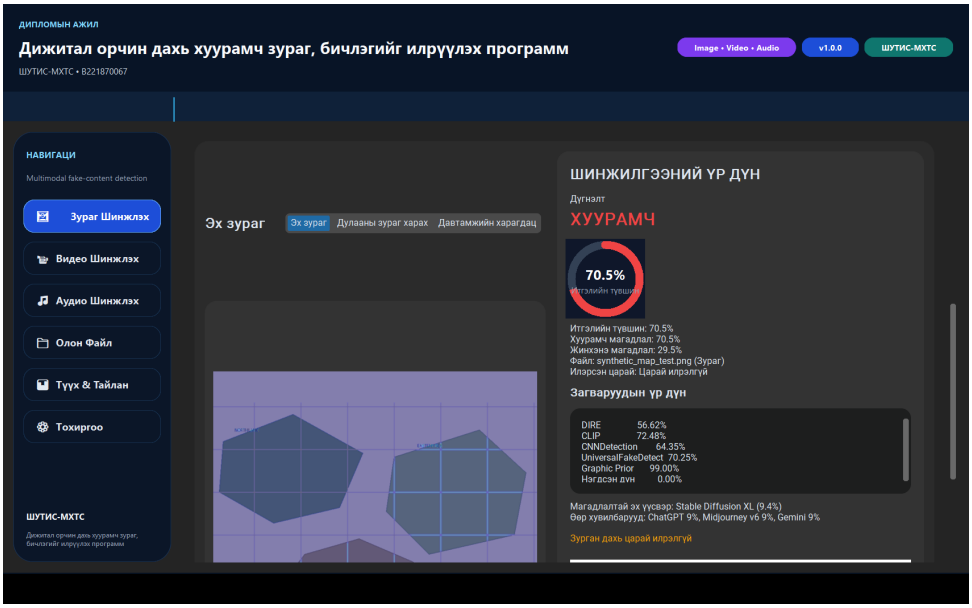
Зураг 3.19: Зураг оруулсан байдал ба нүүр таних явц.

Зураг 3.19-д хэрэглэгч зураг оруулсан байдал болон нүүр таних явцын интерфейсийг харуулав.



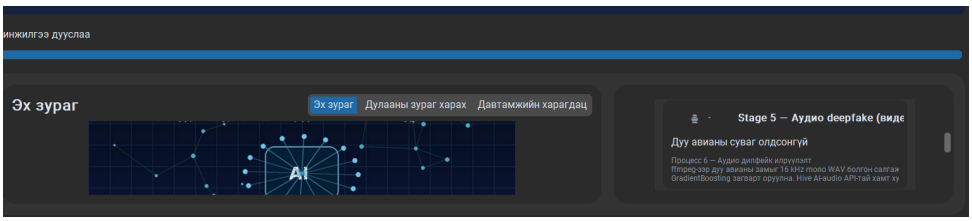
Зураг 3.20: Зураг шинжилгээний эцсийн үр дүн ба үе шаттай дүгнэлт.

Зураг 3.20-д эцсийн дүгнэлт (95.0% итгэлийн түвшинтэй ХУУРАМЧ) болон Шат 0–6 (гарал үүсэл, дипфейк, хиймэл оюун ухаанаар үүссэн, identity, Hive moderation, Celebrity recognition)-ын үе шатуудын задаргаа харагдаж байна.



Зураг 3.21: Загварын сэжигтэй бүсийг дүрслэх дулааны зураглал

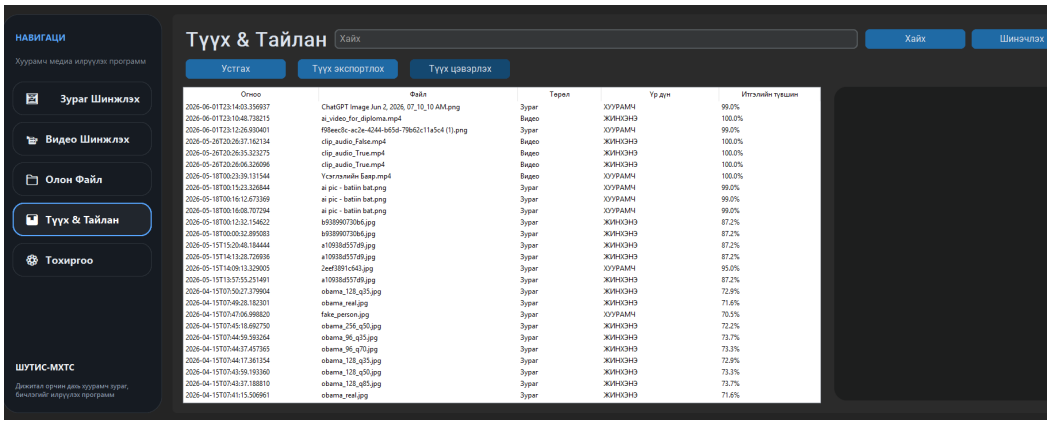
Зураг 3.21-д зургийн илрүүлэлтийн хамт харагдах Grad-CAM дулааны зураглалыг харуулав. Энэ харагдац нь загвар зургийн аль хэсгийг хуурамч гэж дүгнэхдээ ихэвчлэн анхаарч буйг харуулах зорилготой бөгөөд хэрэглэгч дүгнэлтийн шалтгааныг ойлгоход тусалдаг.



Зураг 3.22: Видео доторх аудио дипфейк шинжилгээний үр дүн

Зураг 3.22-д Шат 5 видео хэсэгт байгаа аудио шинжилгээний үр дүнг харуулав. болон Hive AI-audio шалгалтын оноо интерфейсэд харагдана.

Шинжилгээний үр дүн тусламжтай өгөгдлийн санд бүртгэгдэж, цаашид түүхээс хайх, шүүх, харьцуулах боломжтой болсон.



Зураг 3.23: Түүх ба тайлангийн интерфейс.

Зураг 3.23-д шинжилгээний түүх ба тайлангийн интерфейсийг харуулав. Хэрэглэгч огноо, файлын нэр, медиа төрөл, эцсийн дүгнэлт болон итгэлийн түвшинг харж, бүртгэл бүрийг шүүх, экспортлох эсвэл устгах боломжтой.

id	timestamp	file_path	file_type	fake_type	confidence	likely_source
111	2026-06-0	C:\Users\e\Downloads\ai_	image	AI_GENER/	0.99	C2PA Content Credentials (verified)
110	2026-06-0	C:\Users\e\Downloads\ai_	video	REAL	1	Хамаарахгүй
109	2026-06-0	C:\Users\e\Downloads\ai_	image	AI_GENER/	0.99	C2PA Content Credentials (verified)
108	2026-05-2	C:\Users\e\AppData\Loca	video	REAL	1	Хамаарахгүй
107	2026-05-2	C:\Users\e\AppData\Loca	video	REAL	1	Хамаарахгүй
106	2026-05-2	C:\Users\e\AppData\Loca	video	REAL	1	Хамаарахгүй
105	2026-05-1	C:\Users\e\Downloads\Yc	video	AI_GENER/	1	Seedance (Hive)
104	2026-05-1	C:\Users\e\Downloads\AI	image	AI_GENER/	0.99	C2PA Content Credentials (verified)
103	2026-05-1	C:\Users\e\Downloads\AI	image	AI_GENER/	0.99	C2PA Content Credentials (verified)
102	2026-05-1	C:\Users\e\Downloads\AI	image	AI_GENER/	0.99	C2PA Content Credentials (verified)
101	2026-05-1	C:\Users\e\OneDrive\Pict	image	REAL	0.871622892	Imagen
100	2026-05-1	C:\Users\e\OneDrive\Pict	image	REAL	0.871622892	Imagen
99	2026-05-1	C:\Users\e\OneDrive\Pict	image	REAL	0.871622892	Imagen
98	2026-05-1	C:\Users\e\OneDrive\Pict	image	REAL	0.871622892	Imagen
97	2026-05-1	C:\Users\e\OneDrive\Pict	image	AI_GENER/	0.949949927	Flux (Hive)
96	2026-05-1	C:\Users\e\OneDrive\Pict	image	REAL	0.871622892	Imagen
95	2026-04-1	C:\Users\e\OneDrive\Des	image	REAL	0.728762364	NanoBanana
94	2026-04-1	C:\Users\e\OneDrive\Des	image	REAL	0.71554964	Imagen
93	2026-04-1	C:\Users\e\OneDrive\Des	image	AI_GENER/	0.705284505	Stable Diffusion XL
92	2026-04-1	C:\Users\e\OneDrive\Des	image	REAL	0.72211282	Imagen
91	2026-04-1	C:\Users\e\OneDrive\Des	image	REAL	0.737138869	Stable Diffusion 2.x
90	2026-04-1	C:\Users\e\OneDrive\Des	image	REAL	0.732605008	NanoBanana
89	2026-04-1	C:\Users\e\OneDrive\Des	image	REAL	0.728762364	NanoBanana

Зураг 3.24: Тайланг CSV бүтцээр гаргах боломжтой интерфейс

Зураг 3.24-д тайланг гаргаж CSV хувилбартай болгож байгааг харуулсан.

3.4 Бүлгийн дүгнэлт

Энэ бүлэгт дипломын ажлын хүрээнд хэрэгжүүлсэн программын техникийн бүтэц, өгөгдлийн сан, зураг, видео болон аудио шинжилгээний үйл явцыг үе шаттай тайлбарлав.

- **Архитектур.** Систем нь интерфэйс, DetectorService, Predictor, загварын давхарга, гадаад API, хадгалалт гэсэн модульчлагдсан бүтэцтэй. Аудио шинжилгээ нь видео шинжилгээний дотор нэгтгэгдсэн.
- **MN-FaceDF (Монгол царайны дипфейк илрүүлэгч).** 1 124 ширхэг 224× 224 нүүрний crop дээр сургасан ResNet-18 загвар тестийн 168 дээж дээр Accuracy 0.845, F1 0.776, ROC AUC 0.865 үзүүлсэн.
- **Аудио дипфейк.** Видеоны дуу авиаг ffmpeg-ээр салгаж Hive AI-generated Audio API-аар үнэлж, хуурамч дуу хоолойг илрүүлдэг.
- **Гадаад датасет.** FaceForensics++, Celeb-DF, DFDC, DeeperForensics-1.0, UADFV дээрх Xception, EffNet-B4, F3Net, SPSL загваруудын AUC-ийг харьцуулж, нэг загварт найдах нь хангалтгүйг харуулав.
- **Зураг шинжилгээ.** Метадата, эх үүсвэрийн мөр, нүүр илрүүлэлт, хиймэл оюун ухаанаар үүссэн ба нүүрэн дипфейк илрүүлэгчид, Hive API-г жинтэйгээр нэгтгэх 8 үе шаттай.
- **Видео шинжилгээ.** Гурван фрейм сонгож тооцооллын ачааллыг бууруулах ба **дуу авианы сувгийг салгаж аудио дипфейк (Шат 5)** шинжлэх 6 үе шаттай.
- **Тайлан.** Бүх үр дүн өгөгдлийн санд хадгалагдаж, CSV хэлбэрээр гаргадаг.

Дүгнэлт

Энэхүү дипломын ажлын хүрээнд дижитал орчин дахь хуурамч зураг, бичлэгээс үүсэх аюулгүй байдлын эрсдэлийг судалж, тэдгээрийг илрүүлэх аргачлал, өгөгдөл бэлтгэл, загварын үнэлгээ болон программын хэрэгжилтийг нэгтгэн боловсруулж хэрэгжүүлэв. дипломын ажлын үр дүнг дараах байдлаар дүгнэж байна.

1. Хуурамч контент, хиймэл медиа, дипфейк гэсэн ойлголтуудыг ялган тодорхойлж, зураг, видео болон олон модаль хэлбэрээр үүсэх хуурамч контентын үндсэн төрлүүдийг судалсан.

2. GAN, авто-кодлогч, диффузийн загвар, трансформерт суурилсан үүсгэгч загваруудын ажиллах зарчмыг авч үзэж, эдгээр технологи нь бодит мэт зураг, бичлэг үүсгэх боломжтой боловч дүрсний бүтэц, гэрэлтүүлэг, давтамжийн тархалт, сэргээн босголтын алдаа зэрэг илрүүлэх боломжтой шинж үлдээдэг болохыг тодорхойлсон.

3. Хуурамч зураг, видео илрүүлэх аргуудыг орон зайн шинж, давтамжийн шинж, сэргээн босголтын ялгаа, хугацааны уялдаа болон нэгдсэн илрүүлэлтийн арга гэж ангилан судалсан. Үүний үндсэн дээр нэг загварт бүрэн найдах бус, олон илрүүлэгчийн оноог нэгтгэх ансамбль зарчмыг программын аргачлалд ашигласан.

4. FaceForensics++, Celeb-DF v2, DFDC, DeeperForensics-1.0 зэрэг олон улсын видео дипфейк өгөгдлийн сангууд болон ForenSynths, GenImage, DIRE зэрэг хиймэл зураг илрүүлэх өгөгдлийн сангуудыг судалгаанд ашиглаж, илрүүлэгч загваруудын ажиллах орчин, үнэлгээний үндсийг бүрдүүлсэн.

5. Монгол царайны өгөгдөл дээр ResNet-18 архитектурт суурилсан MN-FaceDF загварыг сургаж, тестийн 168 дээж дээр үнэлсэн. Туршилтын үр дүнд Accuracy = 0.845, Balanced Accuracy = 0.818, Precision = 0.865, Recall = 0.703, F1-score = 0.776, Macro F1 = 0.829, ROC AUC = 0.865 үзүүлэлт гарсан.

6. Зураг болон видео шинжлэх desktop application хэрэгжүүлж, хэрэглэгчийн интерфэйс, нүүр илрүүлэх хэсэг, урьдчилсан боловсруулалт, загварын давхарга, Hive AI API холболт, үр дүн хадгалах болон тайлан экспортлох модулиудыг нэг системд нэгтгэсэн.

7. Зураг шинжилгээний үед метаданс, нүүр илрүүлэлт, хиймэл зураг илрүүлэгчид болон нүүрэн дипфейк илрүүлэгчдийн оноог нэгтгэн эцсийн үр дүн гаргадаг болгосон. Видео шинжилгээний үед фрейм сонгох, нүүр тайрах, фрейм бүрийн оноо тооцох, видео түвшний дүгнэлт гаргах дарааллыг хэрэгжүүлсэн.

Иймд энэхүү дипломын ажил нь дижитал орчин дахь хуурамч зураг, бичлэгийг олон загварын оноонд тулгуурлан шинжилж, үр дүнг хэрэглэгчид ойлгомжтой байдлаар харуулах, хадгалах, тайлан гаргах боломжтой программын шийдэл боловсруулснаараа шинэлэг болсон.

Хавсралт А. Эх код

Энэхүү хавсралтад дижитал орчин дахь хуурамч зураг, видео болон аудио контентыг илрүүлэх дипломын ажлын программын хүрээнд ашигласан Python эх кодын гол хэсгүүдийг оруулав. Код нь өгөгдөл унших, нүүр илрүүлэх, гүн сургалтын загвар ашиглан таамаглал хийх, онлайн API-тай холбогдох, видео фрейм задлах, аудио шинж тэмдэг тооцоолох, эцсийн оноо нэгтгэх болон шинжилгээний үр дүнг өгөгдлийн санд хадгалах үндсэн модулиудаас бүрдэнэ.

1 Сангууд болон ерөнхий тохиргоо

Эх код 1-д программын үндсэн модулиудад ашигласан Python сан, гүн сургалтын хүрээ болон ерөнхий тогтмол хувьсагчуудыг харуулав.

Эх код 1: Сангууд болон ерөнхий тохиргоо

```python
from __future__ import annotations

import json
import joblib
from pathlib import Path
from dataclasses import dataclass, field
from typing import Any, Callable

import numpy as np
from PIL import Image

import torch
from torch import nn
from torchvision import models, transforms

from facenet_pytorch import MTCNN, InceptionResnetV1

from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.model_selection import GroupKFold
from sklearn.metrics import (
    accuracy_score,
    precision_score,
    recall_score,
    f1_score,
    roc_auc_score,
    confusion_matrix,
)

RANDOM_STATE  = 42
PROJECT_ROOT  = Path(__file__).resolve().parent
OUTPUT_DIR    = PROJECT_ROOT / "outputs"
CHECKPOINT_DIR = PROJECT_ROOT / "checkpoints"
DEEPFAKE_THRESHOLD = 0.60
AI_IMAGE_THRESHOLD = 0.65
AUDIO_THRESHOLD    = 0.62
```

2 MN-FaceDF загварын сургалт

Эх код 2-д MN-FaceDF өгөгдлийн багц дээр ResNet-18 загварыг сургах үндсэн хэсгийг харуулав. ImageNet-1K-д урьдчилан сургасан жинг ачааллаж, ангиллын давхаргыг бодит/хуурамч хоёр ангилалтай болгосон.

Эх код 2: MN-FaceDF загварын сургалт (ResNet-18, AdamW, 20 давталт)

```
# tools/build_mongolian_image_dataset.py
train_tfm = transforms.Compose([
    transforms.Resize(256),
    transforms.RandomResizedCrop(224, scale=(0.7, 1.0)),
    transforms.RandomHorizontalFlip(),
    transforms.ToTensor(),
    transforms.Normalize([0.485, 0.456, 0.406], [0.229, 0.224, 0.225]),
])

model = models.resnet18(weights=models.ResNet18_Weights.IMAGENET1K_V1)
model.fc = nn.Linear(model.fc.in_features, 2)
model.to(device)

optimizer = torch.optim.AdamW(model.parameters(), lr=args.lr, weight_decay=1e
    -4)
criterion = nn.CrossEntropyLoss()

best_f1 = -1.0
for epoch in range(1, args.epochs + 1):
    model.train()
    for imgs, labels in train_loader:
        imgs, labels = imgs.to(device), labels.to(device)
        optimizer.zero_grad()
        loss = criterion(model(imgs), labels)
        loss.backward()
        optimizer.step()
    metrics = evaluate(model, val_loader, device)
    if metrics["f1"] > best_f1:
        best_f1 = metrics["f1"]
        torch.save({"model": model.state_dict()}, CHECKPOINT_DIR / "best.pt")
```

3 Зургийн гарал үүслийн мэдээлэл задлах

Эх код 3-д зургийн SHA-256 хэш, EXIF/XMP метадата, C2PA баталгаажуулалт болон SynthID усан тэмдгийг нэгтгэн задлах хэсгийг харуулав.

Эх код 3: EXIF / C2PA / SynthID гарал үүслийн мэдээлэл задлах

```
# core/credential_extractor.py
def extract(self, image_path: Path) -> Credentials:
    sha256 = hash_file(image_path, algorithm="sha256")
    exif   = read_exif_xmp(image_path)
    c2pa   = scan_c2pa_manifest(image_path)
    synth  = detect_synthid_watermark(image_path)
    return Credentials(sha256=sha256, exif=exif, c2pa=c2pa, synthid=synth)
```

4 MTCNN ашиглан нүүр илрүүлэх

Эх код 4-д MTCNN загвар ашиглан нүүрийг илрүүлж, итгэлцлийн босго утгаар шүүж, 18%-ийн margin нэмж 224×224 хэмжээтэй тайрах хэсгийг харуулав.

Эх код 4: MTCNN ашиглан нүүрийг илрүүлж тайрах (224x224)

```
# core/face_extractor.py
def detect_faces(self, image):
    boxes, probs = self.mtcnn.detect(image)
    faces = []
    for box, prob in zip(boxes or [], probs or []):
        if prob < self.confidence_threshold:
            continue
        crop = crop_with_margin(image, box, margin=0.18, size=224)
        faces.append(Face(bbox=box, image=crop, probability=prob))
    return faces
```

5 Химийэл оюун ухаанаар үүссэн зураг болон нүүрэн дипфейк ансамбль

Эх код 5-д хиймэл оюунаар үүсгэсэн зураг илрүүлэгч (CLIP, DIRE, CNN, UFD) болон нүүрэн дипфейк илрүүлэгч (Xception, EfficientNet-B4, F3Net, SPSL, MN-FaceDF) загваруудын жинлэсэн дундаж тооцооллыг харуулав.

Эх код 5: AI-зураг ба нүүрэн дипфейк ансамбль (жинлэсэн дундаж)

```
# core/predictor.py
def run_ai_image_ensemble(self, image):
    breakdown = {}
    for name in self.active_ai_models():        # CLIP, DIRE, CNN, UFD
        breakdown[name] = self.model_loader.get(name)\
            .predict_image(image)["fake_score"]
    return weighted_average(breakdown, self.weights["ai_image"]), breakdown

def run_deepfake_ensemble(self, faces):
    breakdown = {}
    for name in self.active_deepfake_models():  # Xception,..., MN-FaceDF
        per_face = [self.model_loader.get(name).predict_image(f)["fake_score"]
                    for f in faces]
        breakdown[name] = sum(per_face) / max(len(per_face), 1)
    return weighted_average(breakdown, self.weights["deepfake"]), breakdown
```

A.6 Hive AI — зураг, видео, аудио онлайн шалгалт

Эх код 6-д Hive AI-ийн v3 endpoint ашиглан зураг, видео болон аудио файлыг шалгах кодод харуулав. Видеог шууд илгээж хариу авна.

Эх код 6: Hive AI онлайн шалгалт (зураг / видео / аудио)

```
# core/hive_detector.py
def analyze(self, image_path: Path | str) -> HiveResult | None:
    if not self.enabled:
        return None
    path    = Path(image_path)
    payload = self._post_multipart(path)
```

```
    if payload is None:
        url = self._upload_to_temporary_host(path)
        if url:
            body    = {"media_metadata": True, "input": [{"media_url": url}]}
            payload = self._post_json(body, label="Hive␣AI")
    return self._parse(payload) if payload else None

def analyze_video(self, video_path):
    result = self.analyze(video_path)
    if result is not None:
        result.media_type = "video"
    return result

def analyze_audio(self, audio_path):
    result = self.analyze(audio_path)
    if result is not None:
        result.media_type = "audio"
    return result
```

7 Зургийн эцсийн оноо нэгтгэх

Эх код 7-д зургийн локал ансамбль болон Hive API-ийн оноог 0.70 : 0.30 жингээр нэгтгэж дүгнэлт гаргах хэсгийг харуулав.

Эх код 7: Зургийн эцсийн оноог нэгтгэж дүгнэлт гаргах

```
# core/detector.py
s_image = self.predictor.combine_scores(ai_breakdown, df_breakdown)
s_hive  = (self.hive_detector.analyze(image_path).fake_probability
           if hive_enabled else 0.0)
s_final = 0.70 * s_image + 0.30 * s_hive

if   s_final >= 0.80: verdict = "FAKE-HIGH"
elif s_final >= 0.60: verdict = "SUSPECT"
elif s_final >= 0.40: verdict = "UNCERTAIN"
else:                 verdict = "REAL"
```

8 Видеоны фрейм бүр дээр нүүрэн дипфейк илрүүлэх

Эх код 8-д видеоноос ялгасан фрейм бүр дээр нүүр илрүүлэгч болон дипфейк ансамбль ажиллах хэсгийг харуулав.

Эх код 8: 100 фрейм бүр дээр нүүрэн дипфейк ансамбль

```
# core/detector.py - analyze_video() loop
for frame in frames:          # 100
    faces = self.face_extractor.detect_faces(frame.image)
    df_score, breakdown = self.predictor.run_deepfake_ensemble(
        [face.image for face in faces])
    tau = self.settings.deepfake_face_threshold   # 0.60
    if   df_score >= tau:        label = "FAKE"; fake_frames += 1
    elif df_score <  tau * 0.85: label = "REAL"
    else:                        label = "UNCERTAIN"
    timeline.append(df_score)
```

.10 Видео доторх аудио замыг Hive AI-аар шалгах

Эх код 9-д видео файлаас дуу авианы замыг ffmpeg-ээр салгаж, Hive AI-generated Audio API-д илгээн хуурамч дуу хоолой эсэхийг үнэлэх хэсгийг харуулав. Дуу авиагүй файлд тусгай төлөв буцаана.

Эх код 9: Видео доторх аудио замыг Hive AI-аар шалгах

```
# core/detector.py
def _analyze_audio_track(self, media_path, progress):
    if not self.hive_detector.enabled:
        return AudioTrackVerdict(has_audio=False)
    try:
        meta, waveform = self.audio_processor.load_audio(media_path)  # ffmpeg
    except Exception:
        return AudioTrackVerdict(has_audio=False)

    suffix = Path(media_path).suffix.lower()
    if suffix in {".mp4", ".avi", ".mov", ".mkv", ".webm"}:
        hr = self.hive_detector.analyze_video(media_path)
    else:
        hr = self.hive_detector.analyze_audio(media_path)
    if hr is None:
        return AudioTrackVerdict(has_audio=False)

    ensemble = float(hr.fake_probability)
    is_fake  = ensemble >= self.settings.audio_fake_threshold
    return AudioTrackVerdict(has_audio=True, ensemble_score=ensemble,
                             is_fake=is_fake, ...)
```

11 Шинжилгээний үр дүнг өгөгдлийн санд хадгалах

Эх код 10-д шинжилгээний огноо, файлын нэр, медиа төрөл, дүгнэлт болон дэлгэрэнгүй payload-ийг `analyses` хүснэгтэд бүртгэх хэсгийг харуулав.

Эх код 10: Шинжилгээний үр дүнг өгөгдлийн санд бүртгэх

```
# utils/history_store.py
def add_result(self, result: DetectionResult) -> int:
    payload = json.dumps(result.to_dict(), ensure_ascii=False)
    with self._connect() as connection:
        cursor = connection.execute(
            """INSERT␣INTO␣analyses
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣(timestamp,␣file_path,␣file_type,␣fake_type,
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣confidence,␣likely_source,␣faces_detected,␣result_json)
␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣VALUES␣(?,␣?,␣?,␣?,␣?,␣?,␣?,␣?)""",
            (result.timestamp, result.file_path, result.file_type,
             result.fake_type, result.confidence, result.likely_source,
             result.faces_detected, payload))
        connection.commit()
        return int(cursor.lastrowid)
```

Хавсралт А. Танилцуулга


ШУТИС
Дижитал орчин дахь хуурамч зураг, видео илрүүлэх системийн хөгжүүлэлт
Development of Fake Images and Videos Detection in Digital Environments
Гүйцэтгэсэн : Ч.Амгалан-Очир /B221870067/
Удирдагч : Х.Уянгаа /Магистр/
Зөвлөгч : Ч.Эрдэнэбат /Доктор(Ph.D)/
Я. Дашдорж /Доктор(Ph.D), дэд профессор/
Кибер аюулгүй байдлын тэнхим
МОНГОЛ УЛСЫН ШИНЖЛЭХ УХААН ТЕХНОЛОГИЙН ИХ СУРГУУЛЬ

Агуулга

I. Сэдвийн үндэслэл
II. Зорилго, Зорилт
III. Хувь нэмэр
IV. Өмнө судлагдсан ажлууд
V. Санал болгож байгаа механизм
VI. Программын архитектур
VII. Туршилтын механизм
VIII. Өгөгдөл багц
IX. Үр дүн
X. Үнэлгээ
XI. Дүгнэлт

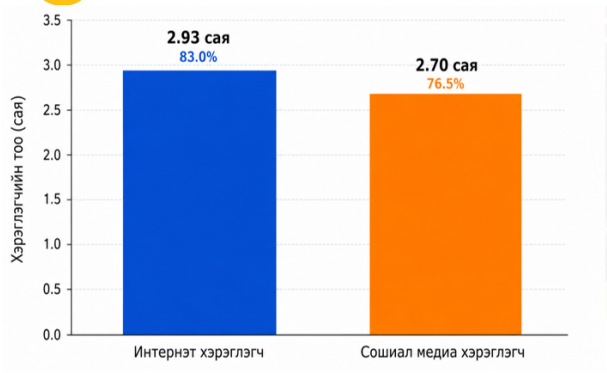
Зорилго, зорилт

Зорилго: Дижитал орчин дахь хуурамч зураг, видеог илрүүлэх програм хангамжийг нээлттэй сан ашиглан хөгжүүлэх.

Зорилтууд:

1. Хуурамч контент, хиймэл медиа, дипфейк ойлголтыг тайлбарлах.
2. Хуурамч зураг, видео үүсгэх технологийн үндсийг судлах.
3. Хуурамч контентын аюулгүй байдлын эрсдэл, нөлөөг тодорхойлох.
4. Илрүүлэх аргачлал, өгөгдлийн сан, түүнтэй холбоотой монгол өгөгдлийн сан үүсгэх , үнэлгээний хэмжүүрүүдээр батлагдсан хэрэгсэл бүтээх
5. Контентын баталгаажуулалт, дижитал мөр судлах.

Сэдэв сонгосон үндэслэл



Зураг 1. Монгол улсын интернет хэрэглэгчийн тоо болон дижитал орчин дахь хэрэглэгийн тоо.[1]

Хүснэгт 1 : Дипфейк контентийн тархалтын 2019-2023 оныг харуулав. [2,3]

Үзүүлэлт	Бодит тоо / өсөлт	Эх сурвалж
Онлайн орчин дахь дипфейк видео	2023 онд **95,820** видео	Security Hero
2019–2023 оны өсөлт	**+550%**	Deloitte, Security Hero
2019 оны суурь тоо	ойролцоогоор **14,678** видео	Deeptrace
Дипфейк ашигласан залилан / таних баталгаажуулалтын халдлага	2022–2023 онд **10 дахин өссөн**	Sumsub

Зураг 2. GAN , Автокодлогч , Диффуз , Трансформер загварыг тайлбарлав.

Судалгаа

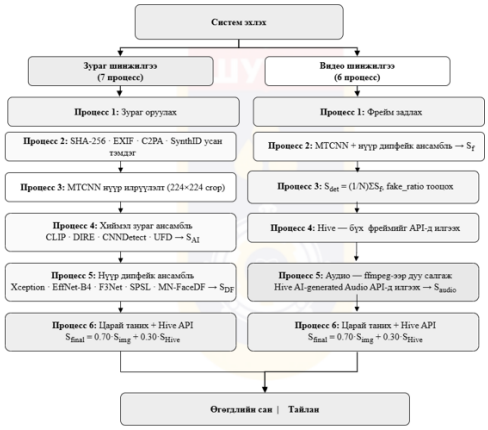
Хүснэгт 3 : Өмнө судлагдсан ажлууд [2,3,4]

Судалгаа	Чиглэл	Гол хувь нэмэр	Энэ ажилд туслах санаа
Chesney & Citron (2019)	Дипфейкий н эрсдэл	Дипфейк нь хувь хүний нууц, итгэлцэл, ардчилал, үндэсний аюулгүй байдалд нөлөөлөх эрсдэлтэйг тайлбарласан.	Асуудлын үндэслэл, эрсдэлийн тайлбар бичихэд ашиглана.
Mirsky & Lee (2021)	Үүсгэх ба илрүүлэх арга	Дипфейк үүсгэх болон илрүүлэх аргуудыг нэгтгэн ангилж, харьцуулсан.	Илрүүлэгчийн аргачлал, detection pipeline тайлбарлахад ашиглана.
Yan et al. (2023)	Туршилт ба үнэлгээ	Олон dataset, detector ашиглан deepfake илрүүлэгч загваруудыг benchmark байдлаар үнэлсэн.	Dataset, baseline model, Accuracy, F1, AUC зэрэг үнэлгээний хэсэгт ашиглана.

Зураг 4. Гадаад өгөгдлийн багцийн тоо хэмжээ

Зураг 5.Хиймэл контентийг илрүүлэх аргуудын график.

Программын архитектур

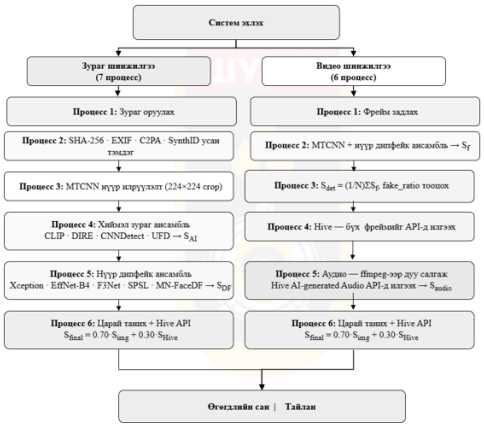


Зураг 6: Программ хангамжийн ерөнхий схем

Программын архитектур

Зураг 7: Программын үндсэн шинжилгээний схем

Программын архитектур



Зураг 6: Программ хангамжийн ерөнхий схем

Туршилтын механизм

Оноо нэгтгэх жин:

Зураг

$S_{final} = 0.70 \cdot S_{image} + 0.30 \cdot S_{HiveAPI}$

Шийдвэрийн босго: ≥0.80 хуурамч магадлал өндөр | 0.60–0.80 сэжигтэй | 0.40–0.60 тодорхойгүй | <0.40 бодит

Видео

S_final = 0.70 · S_video + 0.30 · S_HiveAPI

Шийдвэрийн босго: ≥0.80 хуурамч магадлал өндөр | 0.60–0.80 сэжигтэй | 0.40–0.60 тодорхойгүй | <0.40 бодит

Видео

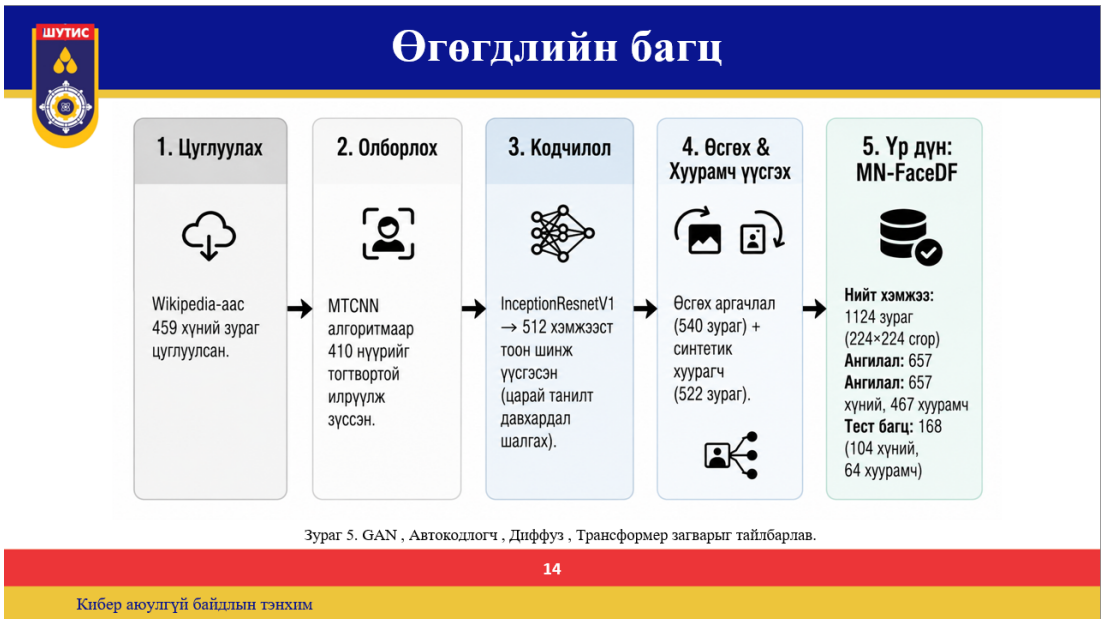
S_audio = S_HiveAPI (зөвхөн Hive AI-generated Audio API)

Шийдвэрийн босго: ≥0.80 хуурамч магадлал өндөр | 0.60–0.80 сэжигтэй | 0.40–0.60 тодорхойгүй | <0.40 бодит

Өгөгдлийн багц

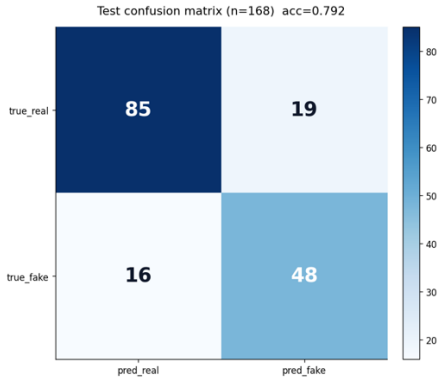
Олон улсын датасет

Загвар	Төрөл	Accuracy	Precision	Recall	F1 оноо	AUC
AI_image / genai_detector	AI зураг	0.944	0.944	0.944	0.944	0.956
dire_detector	AI зураг	0.972	0.972	0.972	0.972	0.996
universal_fake_detector	AI зураг	0.814	0.814	0.814	0.814	0.934
Xception	Дипфейк	0.643	0.615	0.762	0.681	0.722
EfficientNet	Дипфейк	0.579	0.556	0.782	0.650	0.641
F3Net	Дипфейк	0.642	0.608	0.797	0.690	0.707
SPSL	Дипфейк	0.675	0.662	0.707	0.685	0.718

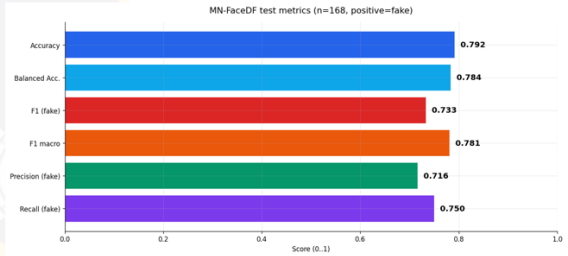


Зураг 5. GAN , Автокодлогч , Диффуз , Трансформер загварыг тайлбарлав.

Өгөгдлийн багц



Зураг 4: 168 өгөгдөл дээр загварын нарийвчлал 79.2% гарч, 133-г зөв, 35-г буруу ангилсан байна..



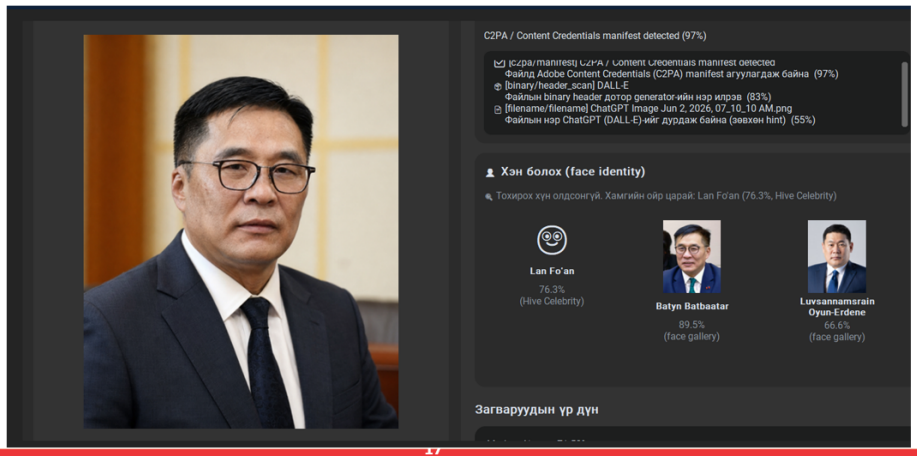
Зураг 5:Сургасан моделийн метрик

Өгөгдлийн багц

Өөрийн датасет

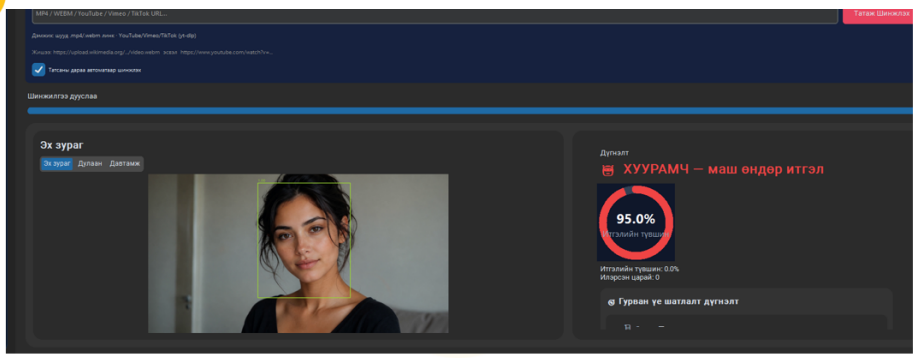
Хэмжүүр	Утга	Тайлбар
Accuracy	0.845	168 тест дүрснээс зөв ангилсан хувь.
Balanced Accuracy	0.818	Жинхэнэ ба хуурамч утгын тэнцвэртэй авч үзсэн утга.
Precision (fake)	0.865	Хуурамч гэсэн нь үнэхээр хуурамч байх хувь.
Recall (fake)	0.703	Бодит хуурамчуудын хэдэн хувийг олж илрүүлсэн.
F1 (fake)	0.776	Precision ба recall-ийн тэнцвэржсэн дундаж.
Macro F1	0.829	Ангилал тус бүрд тооцоод дундажилсан.
ROC AUC	0.865	Босго өөрчлөгдөхөд ялгах чадвар.

Үр дүн



Кибер аюулгүй байдлын тэнхим

Үр дүн



Зураг 10:Программын түүх & тайлан хэсэг

Ерөнхий дүгнэлт

- Дижитал орчин дахь хуурамч зураг, бичлэгээс үүсэх аюулгүй байдлын эрсдэлийг тодорхойлж, олон улсын болон дотоодын түвшинд тулгамдаж буй асуудлыг гаргаж ирсэн.
- Монгол хүний нүүр царайны онцлогийг тусгасан 1124 зураг бүхий “MN-FaceDF” датасетийг амжилттай үүсгэж, илрүүлэлтийн модел бэлтгэсэн.
- Олон AI загварын ансамбль оноо болон Hive API-д тулгуурлан, хэрэглэгчид шууд ашиглах боломжтой илрүүлэлтийн системийг хөгжүүлэв.

Ном зүй

[1] DataReportal (2025). *Digital 2026: Mongolia*. DataReportal. Холбоос: https://datareportal.com/reports/digital-2026-mongolia (Сүүлд хандсан: 2026-05-18)

[2] Deloitte (2025). *The rise of deepfakes: What digital platforms and technology organizations should know*. Deloitte. Холбоос: https://www.deloitte.com/uk/en/Industries/tmt/analysis/the-rise-of-deepfakes-what-digital-platforms-and-technology-organizations-should-know.html (Сүүлд хандсан: 2026-05-18).

[3] Sumsub (2023). *Global Deepfake Incidents Surge Tenfold from 2022 to 2023*. Sumsub. Холбоос: https://sumsub.com/newsroom/sumsub-research-global-deepfake-incidents-surge-tenfold-from-2022-to-2023/ (Сүүлд хандсан: 2026-05-18).

[4]Монгол Улсын Их Хурал (2021). *Хүний хувийн мэдээлэл хамгаалах тухай хууль*. Legalinfo.mn. Холбоос: https://legalinfo.mn/mn/detail?lawId=16390288615991 (Сүүлд хандсан: 2026-05-18)

[5] Eguur.mn (2026). *Хиймэл оюун ухаан, дипфейк технологийн эрсдэлтэй холбоотой мэдээ*. Eguur.mn. Холбоос: https://eguur.mn/ (Сүүлд хандсан: 2026-05-18)


ШУТИС
Анхаарал хандуулсанд баярлалаа.
21
Кибер аюулгүй байдлын тэнхим

Ном зүй

[1] Europol Innovation Lab. *Facing Reality? Law Enforcement and the Challenge of Deepfakes*. 2022. URL: https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-deepfakes (**urlseen** 14/05/2026).

[2] DataReportal. *Digital 2026: Mongolia*. 2025. URL: https://datareportal.com/reports/digital-2026-mongolia (**urlseen** 14/05/2026).

[3] Andreas Rössler **andothers**. "FaceForensics++: Learning to Detect Manipulated Facial Images". **in***Proceedings of the IEEE/CVF International Conference on Computer Vision*: 2019.

[4] Yuezun Li **andothers**. "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics". **in***Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*: 2020, **pages** 3204–3213. DOI: 10.1109/CVPR42600.2020.00327.

[5] Sheng-Yu Wang **andothers**. "CNN-Generated Images Are Surprisingly Easy to Spot... for Now". **in***Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 2020, **pages** 8695–8704. URL: https://arxiv.org/abs/1912.11035.

[6] Zhendong Wang **andothers**. "DIRE for Diffusion-Generated Image Detection". **in***Proceedings of the IEEE/CVF International Conference on Computer Vision*: 2023. URL: https://arxiv.org/abs/2303.09295.

[7] Mingjian Zhu **andothers**. "GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image". **in***Advances in Neural Information Processing Systems 36: NeurIPS 2023 Datasets and Benchmarks Track*: 2023. DOI: 10.48550/arXiv.2306.08571. URL: https://openreview.net/forum?id=GF84C0z45H.

[8] Yisroel Mirsky **and** Wenke Lee. "The Creation and Detection of Deepfakes: A Survey". **in***ACM Computing Surveys*: 54.1 (2021), **pages** 1–41. DOI: 10.1145/3425780.

[9] K. R. Prajwal **andothers**. "A Lip Sync Expert Is All You Need for Speech to Lip Generation in the Wild". **in***Proceedings of the 28th ACM International Conference on Multimedia*: 2020, **pages** 484–492. URL: https://arxiv.org/abs/2008.10010.

[10] Yifan Hu **andothers**. "MnTTS: An Open-Source Mongolian Text-to-Speech Synthesis Dataset and Accompanied Baseline". **in***Proceedings of the 2022 International Conference on Asian Language Processing (IALP)*: 2022, **pages** 184–189. DOI: 10.1109/IALP57159.2022.9961271. URL: https://arxiv.org/abs/2209.10848.

[11] Robin Rombach **andothers**. "High-Resolution Image Synthesis with Latent Diffusion Models". **in***Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 2022, **pages** 10684–10695. URL: https://arxiv.org/abs/2112.10752.

[12] Robert Chesney **and** Danielle Keats Citron. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security". **in***California Law Review*: 107.6 (2019), **pages** 1753–1819. DOI: 10.2139/ssrn.3213954.

[13] The Dalí Museum. *Dalí Lives (via Artificial Intelligence)*. https://thedali.org/exhibit/dali-lives/. Accessed: 2026-05-16. 2019.

[14] Ian J. Goodfellow **andothers**. "Generative Adversarial Nets". **in***Advances in Neural Information Processing Systems*: 2014. URL: https://arxiv.org/abs/1406.2661.

[15] Jonathan Ho, Ajay Jain **and** Pieter Abbeel. "Denoising Diffusion Probabilistic Models". **in***Advances in Neural Information Processing Systems*: 2020. URL: https://arxiv.org/abs/2006.11239.

[16] The Guardian. *UK engineering firm Arup falls victim to £20m deepfake scam*. Accessed: 2026-05-18. 2024. URL: https://www.theguardian.com/technology/article/2024/may/17/uk-engineering-arup-deepfake-scam-hong-kong-ai-video.

[17] World Economic Forum. *Cybercrime: Lessons learned from a* $25m deepfake attack. Accessed: 2026-05-18. 2025. URL: https://www.weforum.org/stories/2025/02/deepfake-ai-cybercrime-arup/.

[18] Mongolian Fact-Checking Center. *Mongolian Fact-Checking Center*. 2026. URL: https://mfcc.mn/ (**urlseen** 14/05/2026).

[19] Deloitte. *The rise of deepfakes: What digital platforms and technology organizations should know*. Deepfake content on social media platforms grew 550% between 2019 and 2023. 2024. URL: https://www.deloitte.com/uk/en/Industries/tmt/analysis/the-rise-of-deepfakes-what-digital-platforms-and-technology-organizations-should-know.html (**urlseen** 16/05/2026).

[20] Zhiyuan Yan **andothers**. "DeepFakeBench: A Comprehensive Benchmark of Deepfake Detection". **in***Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*: 2023. URL: https://arxiv.org/abs/2307.01426.

[21] Coalition for Content Provenance and Authenticity. *C2PA Technical Specification*. Official specification. 2024. URL: https://spec.c2pa.org/specifications/specifications/2.4/ (**urlseen** 14/05/2026).

[22] Hive AI. *AI-Generated and Deepfake Content Detection API*. 2026. URL: https://thehive.ai/apis/ai-generated-content-classification (**urlseen** 14/05/2026).