Монгол Улсын Их Сургуулийн
Хэрэглээний Шинжлэх Ухаан,
Инженерчлэлийн Сургууль

Машин сургалтын хэрэглээ: бичвэр, зургийн шинжилгээ

Док. Д.Энхзол

Мэдээлэл, компьютерийн ухааны тэнхимийн
ахлах багш

2022.03.23

Агуулга

- Тэнхимийн танилцуулга

- Машин сургалтын талаар товч
- 1. Бичвэр дээрх шинжилгээ (ангилал)
- 2. Машин сургалт: регрессийн таамаглал
- 3. Гүн сургалт: зурган өгөгдөл

МКУТ: Нийт 52 багш, ажилтантай

Доктор багш: 25
(профессор, дэд профессор, ахлах багш, багш)

Магистр багш: 19

Лаборант: 6
Тэнхимийн туслах: 2

Монгол, Япон, БНСУ, АНУ, ОХУ, БНХАУ, Итали, Португал, Тайван зэрэг улсуудад боловсролын зэрэг эзэмшиж, ажиллаж байсан.

Судалгааны чиглэл

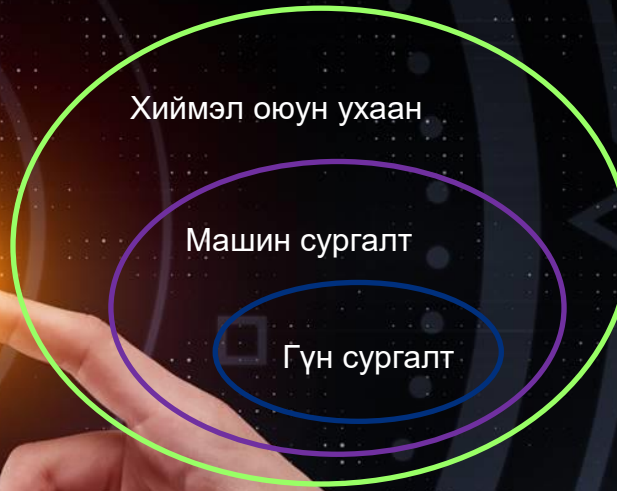
- Програм хангамжийн инженерчлэл
- Хиймэл оюун ухаан
- Мэдээллийн аюулгүй байдал, хамгаалалт
- Өгөгдлийн уурхай, их хэмжээний өгөгдөл боловсруулалт
- Газарзүйн мэдээллийн систем
- Эх хэл боловсруулалт
- Цахим засаглал
- Цахим арилжаа
- Дүрс боловсруулалт зэрэг

Хөтөлбөрүүд

- Бакалаврын:
  - Компьютерийн ухаан
  - Програм хангамж
  - Мэдээллийн технологи
  - Мэдээллийн систем
- Магистр, докторын хөтөлбөрүүд

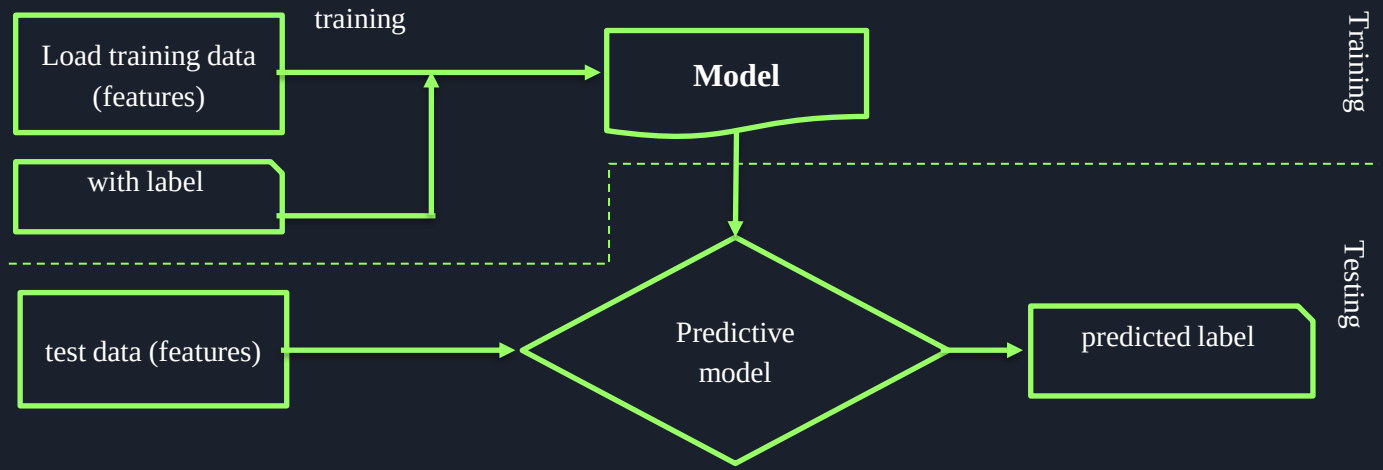
Машин сургалт

Машин сургалт нь компьютерийг шууд програмчлалгүйгээр өмнөх туршлагаасаа (өгөгдлөөс) суралцан, таамаглал дэвшүүлэх боломжтой болгож буй компьютерийн ухааны салбар юм.



- Text analysis
- Security: spam filtering
- Credit card fraud detection
- Digit recognition on checks, zip codes
- Detecting faces in images
- MRI image analysis
- Recommendation system
- Search engines
- Handwriting recognition
- Scene classification
- etc...

Supervised machine learning workflow



Өгөгдлийн төрөл

Text Text Text Text Text

_x005F_x005F_x005F_x005F_x0002_ Numbers

_x005F_x005F_x005F_x005F_x0002_ Clickstreams

_x005F_x005F_x005F_x005F_x0002_ Graphs

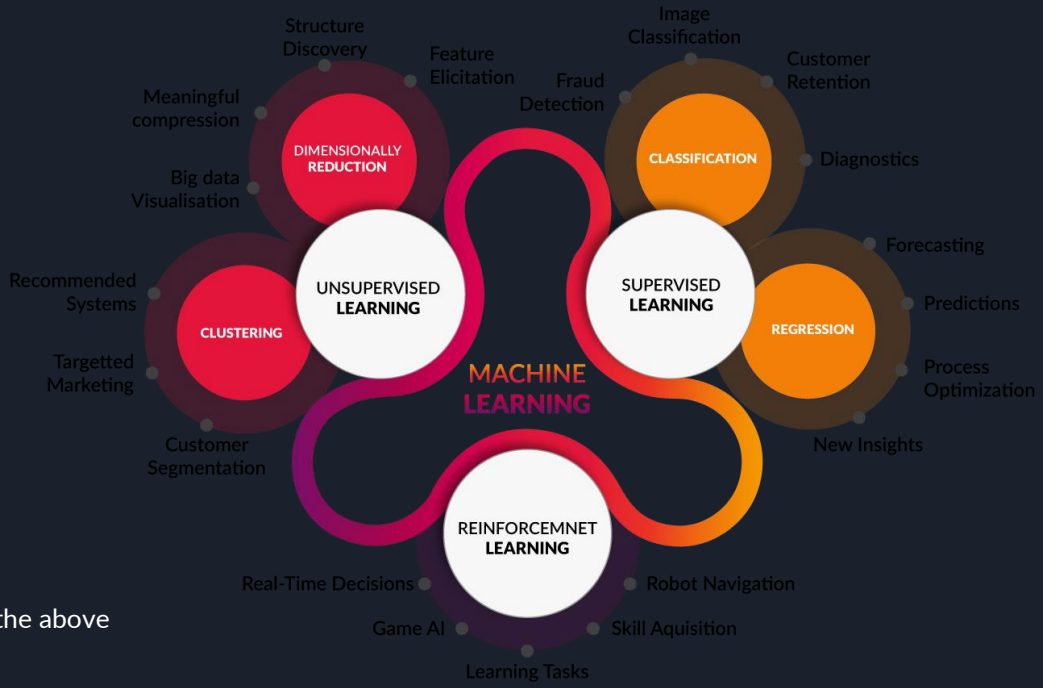
_x005F_x005F_x005F_x005F_x0002_ Tables

_x005F_x005F_x005F_x005F_x0002_ Images

_x005F_x005F_x005F_x005F_x0002_ Transactions

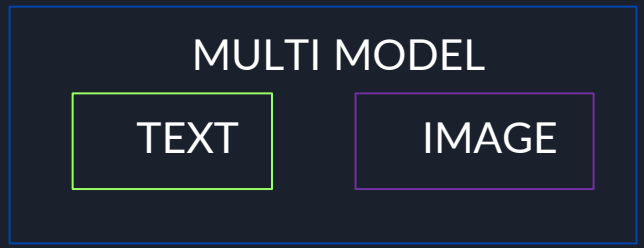
_x005F_x005F_x005F_x005F_x0002_ Videos

_x005F_x005F_x005F_x005F_x0002_ Some or all of the above

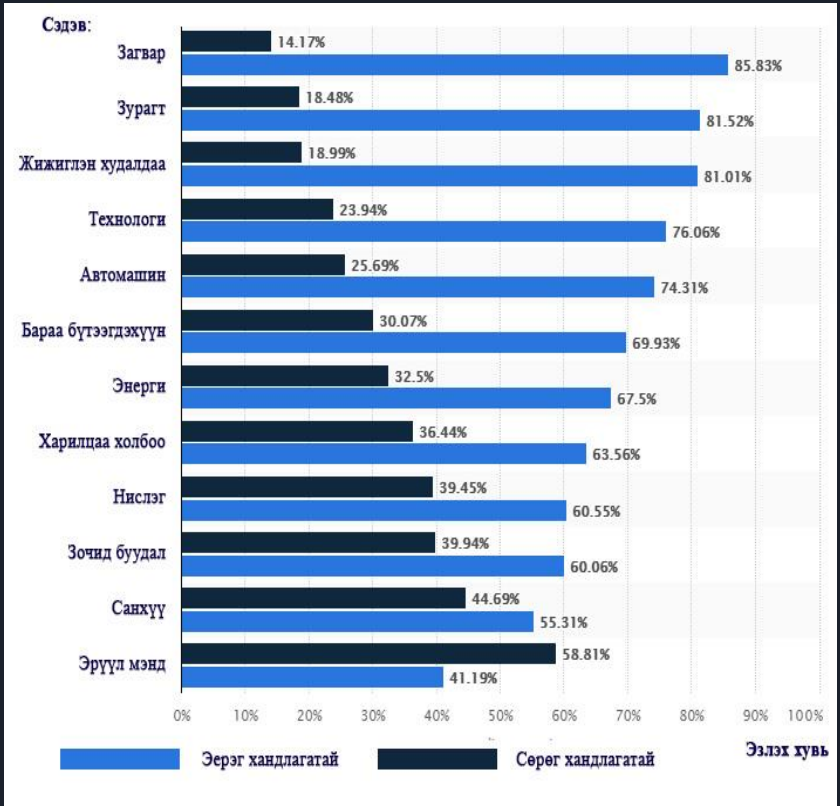


1. Бичвэр дээрх шинжилгээ (ангилал)

- Бичвэрээс өнгөт зураг үүсгэх шинэ аргыг боловсруулж, жиргээний ерөнхий хандлагыг ангилан шинжилж, үнэлэх зорилго тавьсан.



Олон нийтийн сүлжээ Твиттерт 1 секунд тутам 6 мянга, өдөрт 500 сая жиргээ нэмэгддэг.

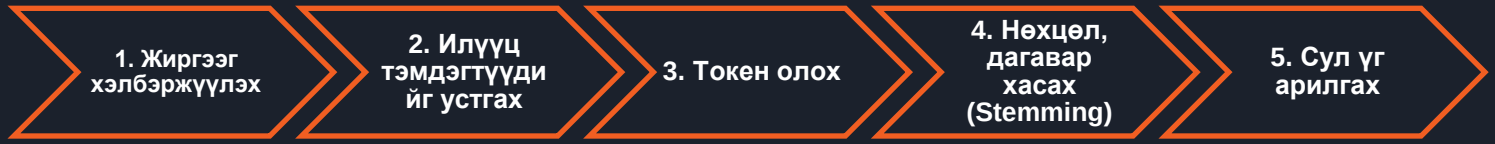


- Тухайн сэдвийн дагуу хэдэн зуу, мянган эерэг, сөрөг, хэвийн хандлагатай жиргээ бичигддэг.
- Хэдэн жиргээ уншаад нийгмийн хандлагыг бүрэн мэдэх бэрх;
- Хандлагыг шинжлэх автомат системийн хэрэгцээ бий болох;
- Бизнесийн болон улс төрийн байгууллага, хувь хүмүүс:
  - тухайн цаг үед,
  - бага зардлаар,
  - богино хугацаанд нийтлэгч, нийгмийн үзэл бодол, хандлагыг мэдэрснээр үйл ажиллагаагаа үнэлэн сайжруулах боломж бүрдэх юм.

Supervised classification: sentiment analysis

- Жиргээний онцлог:
  - Твиттерд 280 тэмдэгтэд багтаан зөв бичих дүрмийн, дүрмийн бус байдлаар, URL, “#, @, emoji” зэрэг хэрэглэн үзэл бодлоо илэрхийлдэг.гээг плот зургаар илэрхийлэ
- I. Жиргээний ерөнхий хандлагыг олох:
  - Жиргээ: “I'm about to eat four hot dogs and watch Miss USA. Happy Sunday. Cheering on @BrookeFletcher http: //t.co/0YKeHN4BABC“
  - Хандлага: ? (үр дүн: эерэг / сөрөг / хэвийн)
- II. Зорилтот сэдвийн дагуу жиргээний хандлагыг олох:
  - Жиргээ (tweet): "Today is Amazon Prime Day. Today is the best day to buy. July 15th only. http://t.co/mfrBPDpL0L #savemoney"
  - Зорилтот сэдэв: “amazon prime day”
  - Хандлага: ? (үр дүн: эерэг / сөрөг / хэвийн)

Нийтлэг боловсруулалтын алхмууд [1], [2]



Боловсруулах алхмууд:

Ашигласан хэл боловсруулалтын сан:

- “Stanford core NLP”, “Open NLTK libriaries”: Tokenization, POS tagger, Lemmatization г.м.

Хэл боловсруулалтын нөөц:

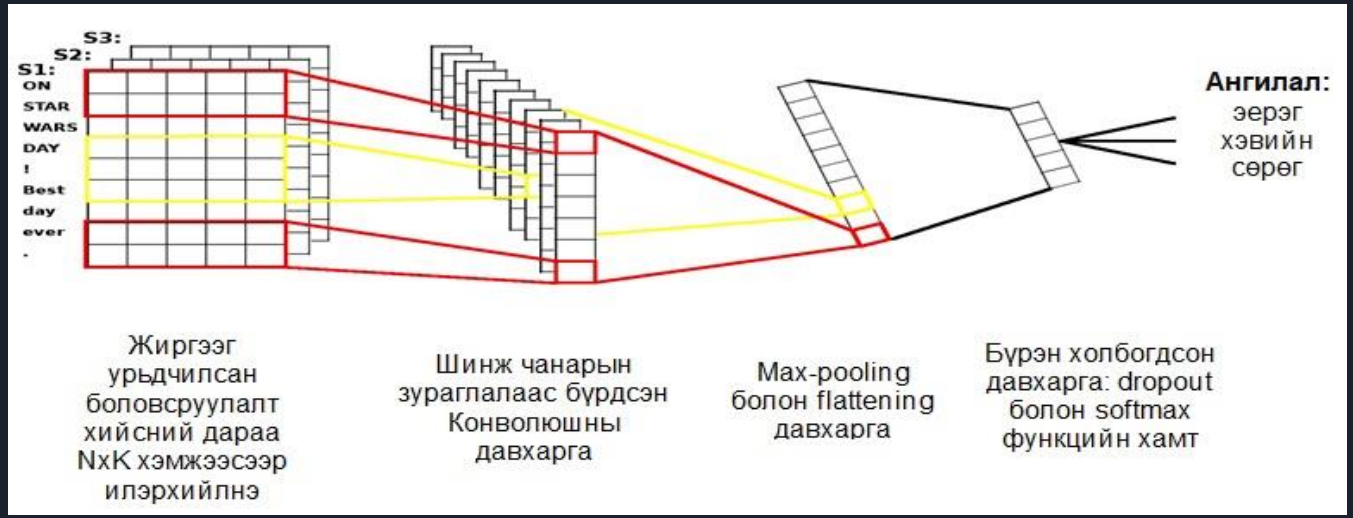
- Эерэг ба сөрөг үгсийн жагсаалт, “SenticNet-5”, “SentiWordNet3.0(117,659 N-грамм үгс)”, “Word2Vec”

[1] B. Duncan and Y. Zhang. Neural networks for sentiment analysis on twitter. … . IEEE, 2015.

[2] H. Saif, M. Fernandez, Y. He, and H. Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. 2014.

Бичвэрийн хэлбэрээс машин, гүн сургалтад ашиглахуйц тоон (скаляр, вектор) хэлбэрт оруулах шаардлагатай.

I. Өргөтгөсөн үгийн шигтгэл: CNN-ийн гүн сургалтын бүтэц



- **“Word2Vec” (Mikolov et al.,2013) загварыг өгөгдлийг 300 хэмжээст векторт хөрвүүлэхэд хэрэглэдэг.**
- **300 хэмжээст “Word2Vec” дээр + “SenticNet5” нөөцөөс *“туйлшрал, тааламжтай байдал, анхаарамж, мэдрэмж, нийцэл”* утгыг нэмж 305 хэмжээст болгож боловсруулсан өгөгдлийг векторт хөрвүүлэхэд хэрэглэж сургасан.**

II.A Жиргээг энгийн вектороор илэрхийлэх суурь арга (baseline)

Жиргээний анхны хувилбар:

I’m about to eat four hot dogs and watch Miss USA. Happy Sunday. Cheering on @BrookeFletcher http://t.co/0YKeHN4BABC

Өмнөх 10 алхамт боловсруулалтын дараа:

“I be about to eat four hot dog and watch Miss USA . happy Sunday . cheer on”

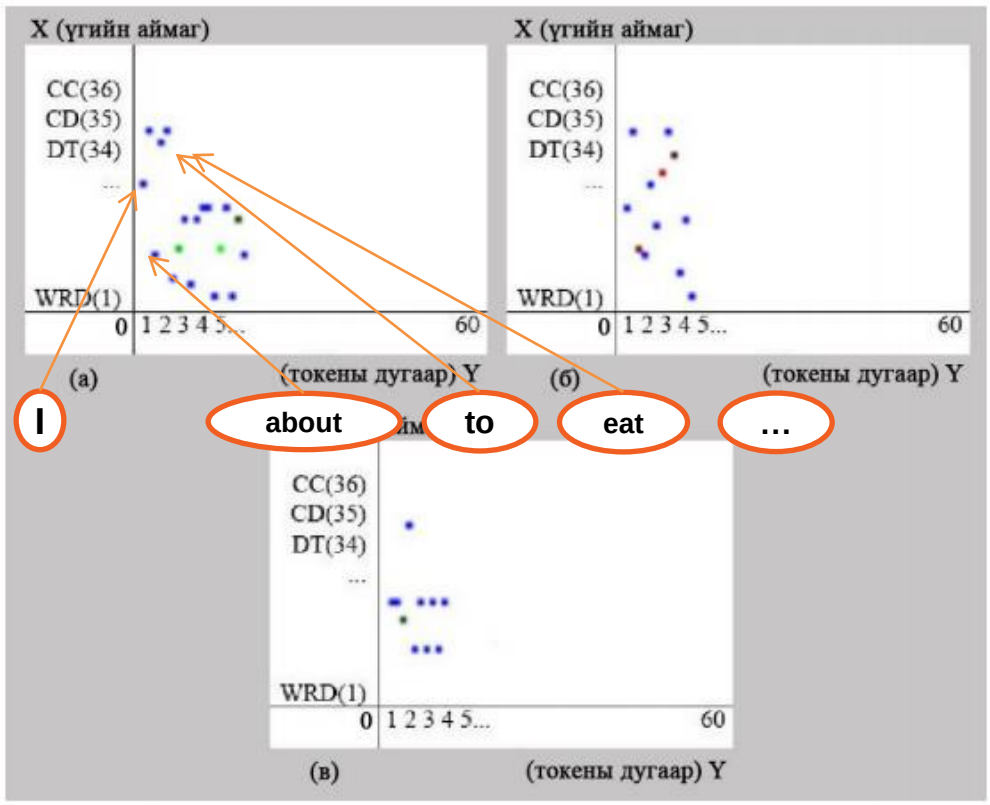
“[0.0,17], [0.0,26], [0.0,5], [0.0,24], [0.0,26], [0.0,1], [0.5,6], [0.0,11], [0.0,11], [0.0,13], [0.0,13], [0.0,0] [0.75,6], [0.0,13], [0.0,0], [0.125,11], [0.0,5], [0.0,0],..., [0.0,0]”

$$embedded_t = [s^{w_1}, p^{w_1}], [s^{w_2}, p^{w_2}], ..., [s^{w_n}, p^{w_n}]$$

Үүнд “hot dog”, “happy”, “cheer” үгнүүд харгалзан [0.5,6], [0.75,6], [0.125,11] мэдрэмжийн утгуудтай бусад үгс мэдрэмжийн хувьд [0.0,*] байна.

II.Б Жиргээг диаграм зургаар илэрхийлэх

Өнгөт цэгээр илэрхийлэгдсэн жиргээний зураг



Жиргээний хандлагаар нь:

- **(a) эерэг**
- **(b) сөрөг**
- **(c) хэвийн**

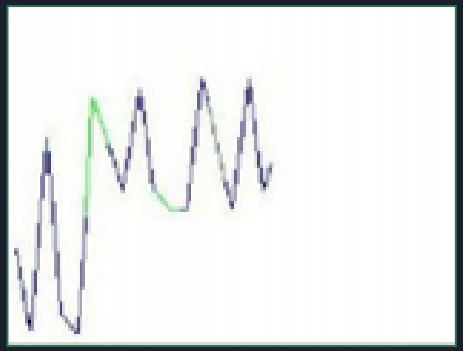
- **Босоо тэнхлэгийн дагуу үгсийн аймаг байршуулсан,**
- **Хэвтээ тэнхлэгийн дагуу жиргээн дэх үгнүүдийг дэс дарааллын дагуу**

Зургийн хэмжээ: 60 x 40

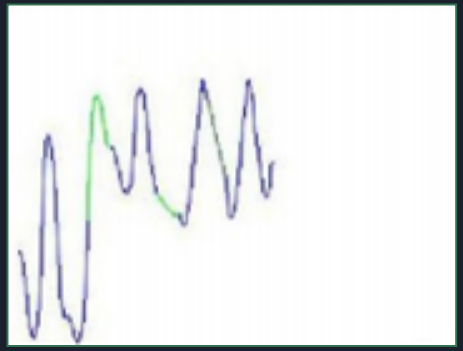
II.Б Жиргээг диаграм зургаар илэрхийлэх

Өнгөт цэгээр илэрхийлэгдсэн жиргээний зураг

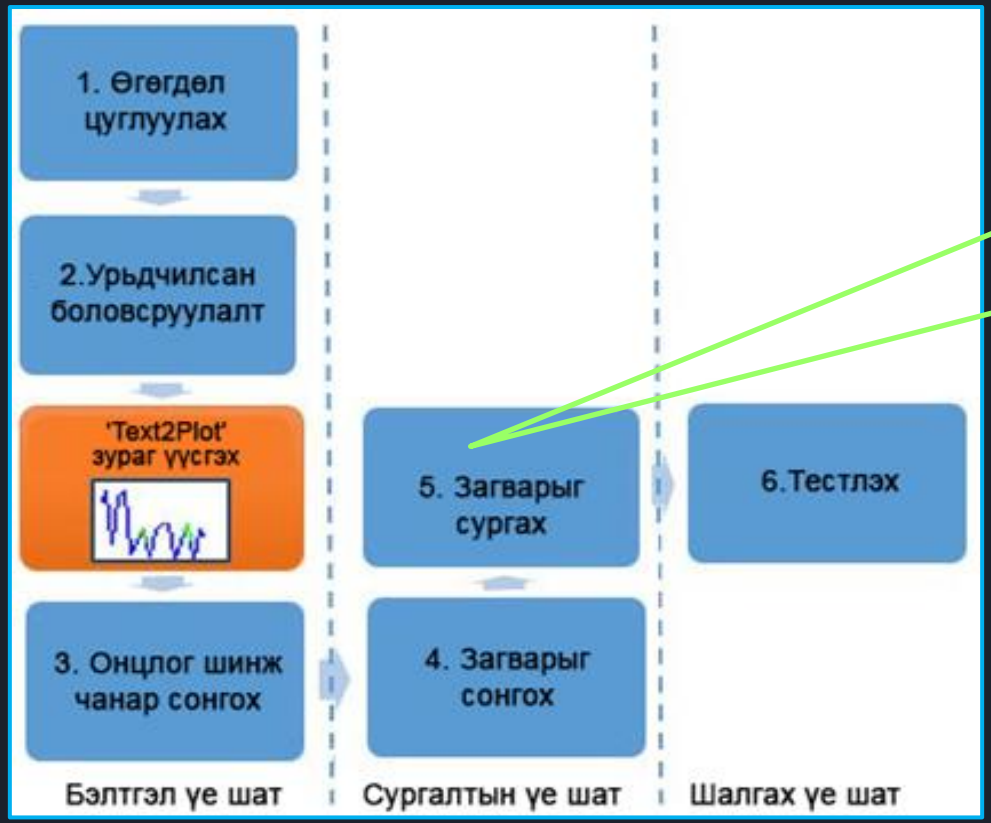
Өнгөт хэрчмүүдээр илэрхийлэгдсэн жиргээний зураг



Интерполяци хэрэглэсэн жиргээний өнгөт зураг



II.Б Жиргээг диаграм зургаар илэрхийлэх



Machine Learning: SVM, Random Forest, Decision Tree, etc.

Deep learning: deep CNN

“Text2Plot” болон үгийн шиггтэлийн аргуудын үр ашигтай байдлын харьцуулалт

Нийтлэг үгийн шиггтэл:

- Word2Vec,
- Glove,
- FastText,
- BERT

Диаграм зургууд :

- Орон зайн нийлмэл байдал (space complexity):
  - 1.36-аас 12.5 дахин бага санах ой зарцуулах
- Хугацааны нийлмэл байдал (time complexity) ердийн үзүүлэлттэй компьютерын хувьд:
  - PNG өргөтгөлтэй зургийн файлыг үүсгэхэд дунджаар 0.2 секунд

Уг дэвшүүлж буй арга нь компьютерын нөөцөд хэмнэлттэй байна.

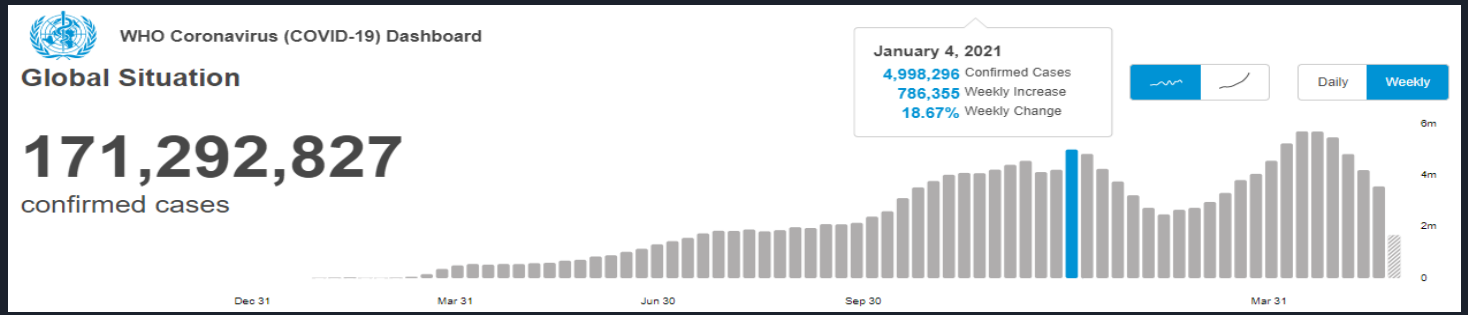
- **Урьдчилсан боловсруулалтад нэмэлт алхмуудыг оруулснаар ангилалын нарийвчлалыг 2% хүртэл нэмэгдүүлж болж байна. Тухайлбал:**
  - Ф1-онооны хувьд URL устгаснаар 0.3%, # таг өөрчилснөөр 1%, @ хэрэглэгчийн нэр өөрчилснөөр 0.4% өссөн ба сургалтын хугацаа 35% хэмнэж хурдассан.
- **Машин болон гүн сургалтад “Text2Plot” аргаар үүсгэсэн өгөгдлийг ашигласнаар ангилалын нарийвчлалыг нэмэгдүүлж байна. Үүнд:**
  - харьцуулсан векторт шилжүүлсэн өгөгдлөөс CNN - 27.2%, SVM - 10.3%, Random Forest - 4.4%-аар илүү ангилсан.

2. Машин сургалт: регрессийн таамаглал

- “COVID-19-ийн дэлхий даяарх болон Монгол Улс дахь тархалтыг Экспоненциаль мөлийлгөлтийн арга ашиглан таамаглах нь” ММТ-2020

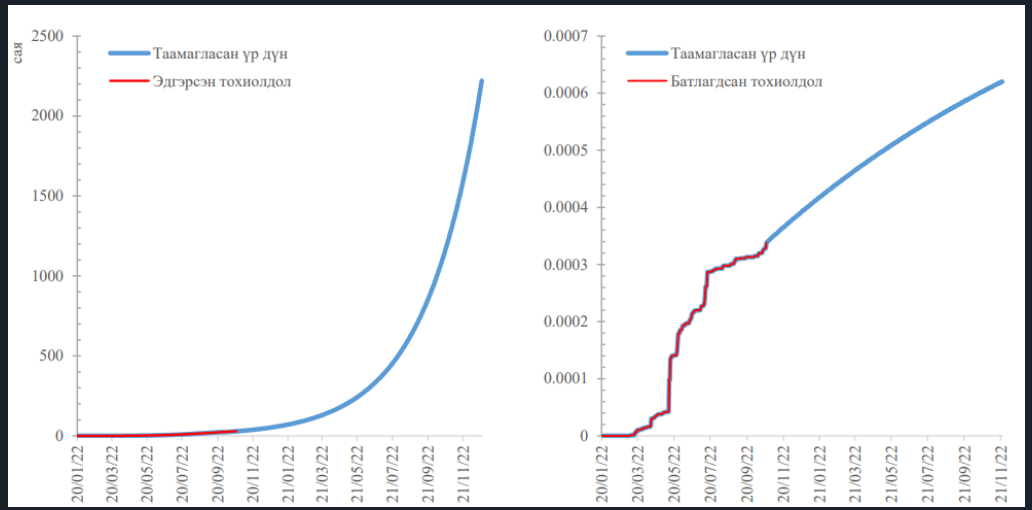
- Forecasting of the COVID-19 Spreading in Global using the Exponential Smoothing Method

- “Монгол Улс дахь цар тахлын тархалтыг Экспоненциаль мөлийлгөлтийн арга ашиглан таамаглах нь“ eNation – 100, 2021



Холтын хэв шинжийн мөлийлгөлтийн аргуудыг сонгон авч машин сургалт хийсэн.

Дэлхий дахины хувьд:



Монгол Улсын хувьд:

- “ММТ-2020” хуралд оролцсон судалгаа үр дүнгээс харахад Дэлхий дахин дахь коронавирусийн тархалтыг 2020.2.20-с 2020.9.30-ны хоорондох нийт 253 өдрийн өгөгдлөөр сургаж ирэх нэг жилийн турш дахь таамагласан.
- Урьдчилсан таамагласан тоон үзүүлэлт нь ~4%-ийн алдаатай буюу ~96% зөв таамагласан байна

Урьдчилсан таамагласан тоо	Бодит батлагдсан тохиолдлын тоо	Зөрүү /тоогоор/	Зөрүү /хувиар/
80,206,954	83,559,601	3,352,647	4,01%

Таамаглал ба урьдчилсан сэргийлэлт

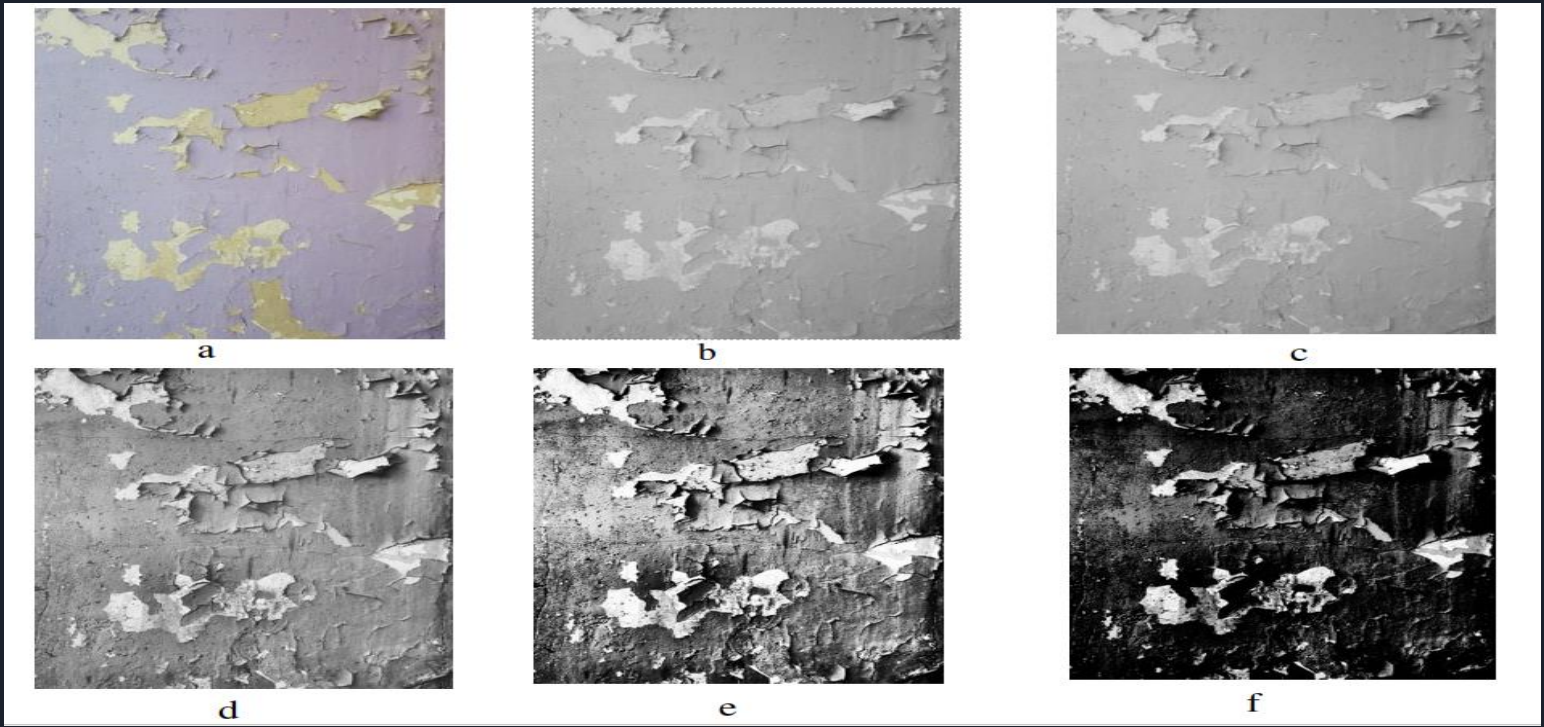
Огноо	Дэлхий нийтэд тархсан хурдаар МУ-д тархсан бол	Дэглэм, вакцинжуулалт 60% хамгаалсан	Дэглэм, вакцинжуулалт 80% хамгаалсан	Дэглэм, вакцинжуулалт 90% хамгаалсан
21/06/12	359,275.27	143,710.11	71,855.05	35,927.53
21/06/13	378,165.64	151,266.25	75,633.13	37,816.56
21/06/14	398,049.24	159,219.70	79,609.85	39,804.92
21/06/15	418,978.32	167,591.33	83,795.66	41,897.83
21/06/16	441,007.82	176,403.13	88,201.56	44,100.78
21/06/17	464,195.61	185,678.24	92,839.12	46,419.56
21/06/18	488,602.60	195,441.04	97,720.52	48,860.26
21/06/19	514,292.88	205,717.15	102,858.58	51,429.29
21/06/20	541,333.94	216,533.58	108,266.79	54,133.39
21/06/21	569,796.79	227,918.72	113,959.36	56,979.68
21/06/22	599,756.19	239,902.48	119,951.24	59,975.62
21/06/23	631,290.83	252,516.33	126,258.17	63,129.08
21/06/24	664,483.54	265,793.41	132,896.71	66,448.35
21/06/25	699,421.48	279,768.59	139,884.30	69,942.15
21/06/26	736,196.43	294,478.57	147,239.29	73,619.64
21/06/27	774,904.98	309,961.99	154,981.00	77,490.50
21/06/28	815,648.79	326,259.51	163,129.76	81,564.88
21/06/29	858,534.86	343,413.95	171,706.97	85,853.49

3. Гүн сургалт: зурган өгөгдөл

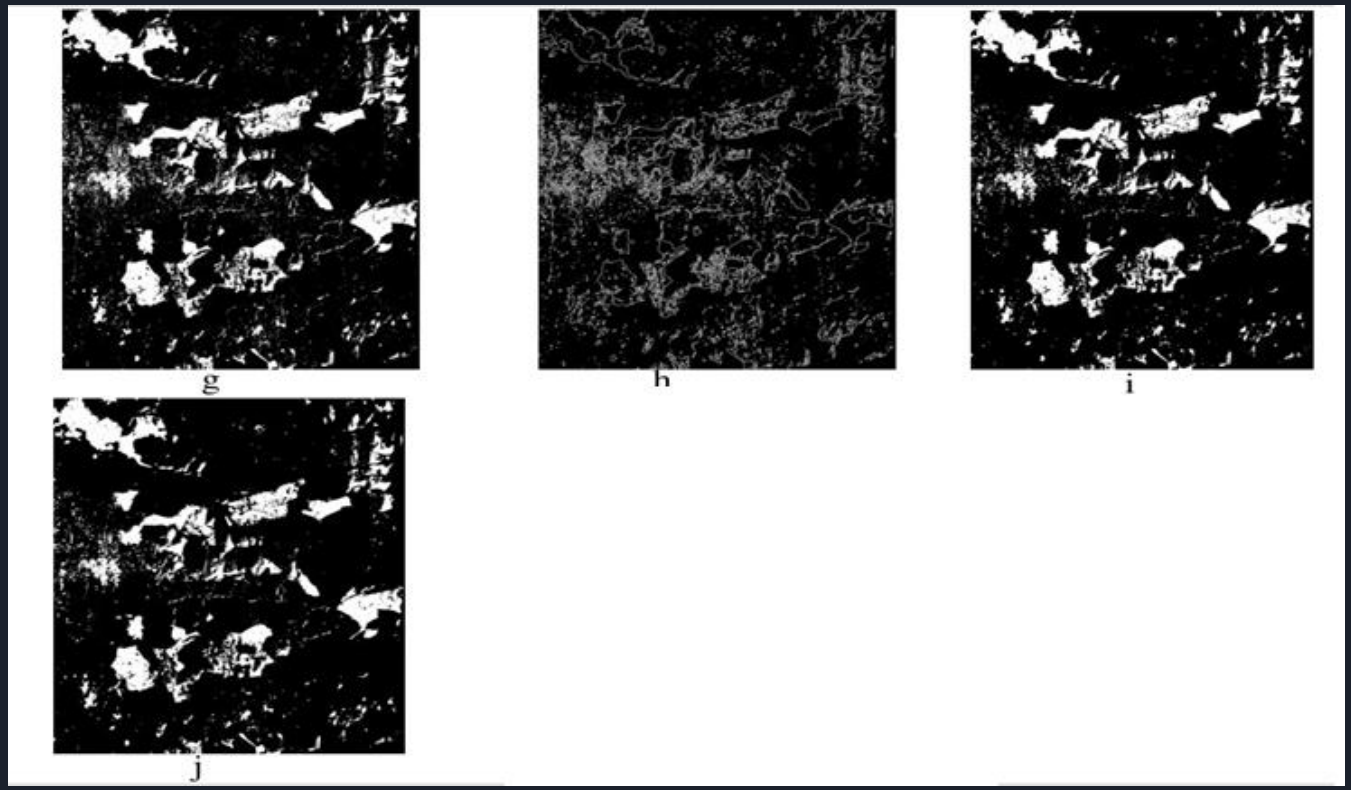
- Одоогийн судалгааны ажил
- **Automatic Detection of Building Surface Scratches :** Image processing, deep learning

  - Барилга, байгууламжийн гадаргуугийн сэв, согогийг зураг боловсруулалт болон машин сургалтын аргуудыг хослуулан илрүүлэх систем бүтээснээр хэрэглэгч бодит хугацаанд үр дүнг харах боломжтой болно.

Preprocessing: image processing

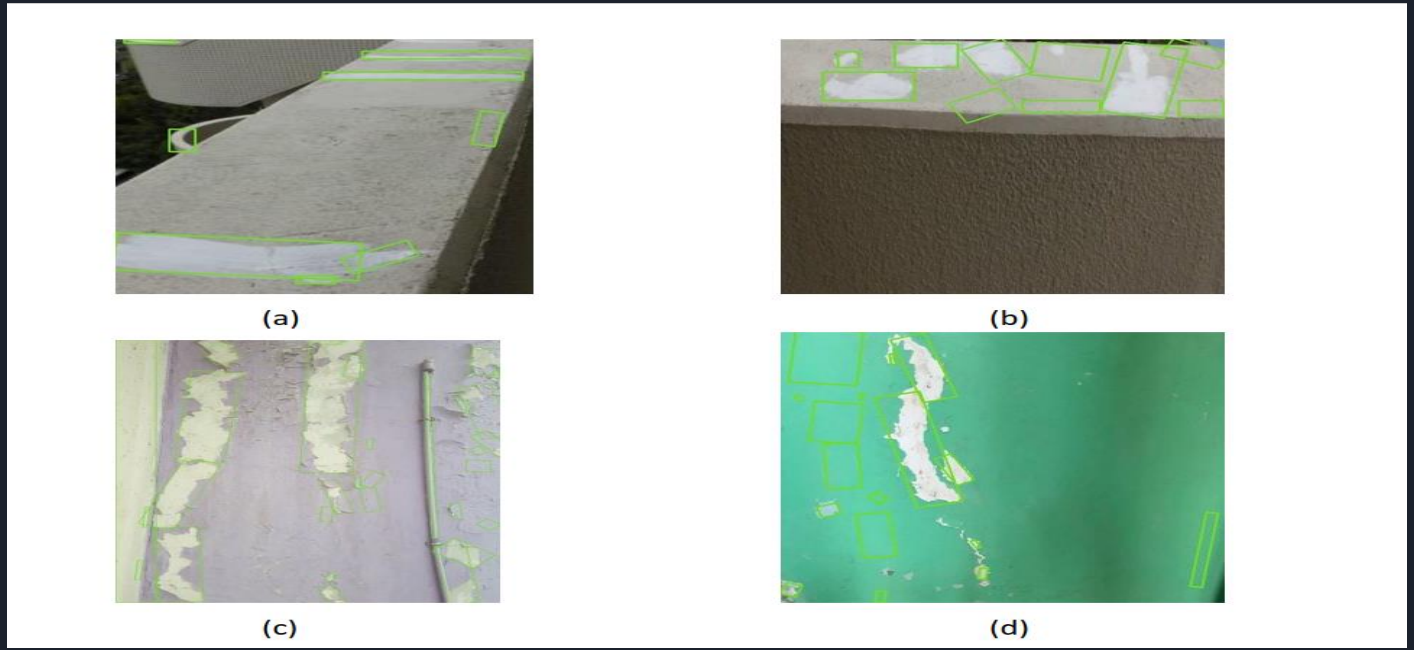


a = Эх зураг , b = Саарал зураг c = Гаусс аргын үр дүн, d = Гистограмм аргын үр дүн, e = Клач тодосгох аргын үр дүн , f = Гамма хувиргалтын үр дүн



g = Оцугийн арга үр дүн , h = Собелын арга үр дүн , i = Морфологийн арга үр дүн

After preprocessing: Faster deep CNN with 10 hidden layer



a = Тавцан дээрх сэв,
b = Цонхны давцан дээрх сэв,
c = Ханан дээрх гадаргуугаас олсон сэв,
d = Өөр хана дээрх гадаргуугаас олсон сэвнүүдийн үр дүн

- Дүрс танилтын олон төрлийн судалгааны суурь болох боломжтой:
  - Барилга орон сууц, автомашин зэргийн засварын үнэлгээг автоматаар гаргах
  - Малын тамга таних
  - Гадны сүг зураг таних гэх мэт.

Анхаарал хандуулсанд баярлалаа ☺