# Монгол хүний цээжний рентген зурагт DenseNet121 загварын шийдвэрийн босго болон магадлалын тохируулгыг үнэлсэн нь

*Д.Баясгалант[1], Э.Оргил-Эрдэнэ[2], Л.Оюунболд[3], Х.Дэлгэрдалай[4]*

*[1]АШУҮИС, АУС, Радиологийн тэнхим*
*[2]AND Global ХХК, Хиймэл оюун ухааны хэлтэс*
*[3]УНТЭ, Радиологийн тасаг*
*[4]АШУҮИС, Радиологийн тэнхим*
*Цахим шуудан: bayasaadc@gmail.com, Утас: 80909363*

**Түлхүүр үг:**
Мэдрэг чанар
Өвөрмөц чанар
Шийдвэрийн босго
Калибраци
Гадаад баталгаажуулалт

**Товч утга:**
**Үндэслэл:** Гадны өгөгдлөөр сургасан загварыг эмнэлзүйн хэрэглээнд нэвтрүүлэхийн өмнө Монгол орны өгөгдөлд ялган таних чадвар, шийдвэрийн босго, тооцсон магадлалын нийцлийг үнэлэх хэрэгцээ байгаа нь энэхүү судалгаа үндэслэл болж байна. **Зорилго:** Олон нийтэд нээлттэй CheXpert өгөгдөл дээр DenseNet121 гүн сургалтын загвар сургаж, түүний гүйцэтгэлийг Монгол хүний цээжний рентген зураг дээр үнэлэн, орон нутгийн өгөгдөлд тохируулах боломжийг судлах. **Арга, аргачлал:** Энэхүү судалгаанд ImageNet дээр урьдчилан сургасан жинг эхлэл болгон DenseNet121 архитектурт суурилсан гүн сургалтын загварыг CheXpert өгөгдөл дээр сургав. Загвар нь 320×320 пикселийн RGB цээжний рентген зургийг оролт болгон авч, эмгэг олдворгүй болон дэмжих хэрэгслийг багтаасан 14 ангилал тус бүрийн илрэх магадлалыг сигмоид функцээр тооцоолов. CheXpert өгөгдлийг сургалт болон загвар сонгох дотоод баталгаажуулалтад 80/20 харьцаатай хуваав. Сонгосон загварыг өмнөх эмнэлзүйн дүгнэлтээс радиологич эмч 14 ангиллын шошгыг гараар гаргасан Монгол хүний 671 цээжний рентген зураг дээр гадаад баталгаажуулалтын хэлбэрээр үнэлэв. Ялган таних чадварыг ROC-AUC, макро AUC болон микро AUC-аар үнэлж, босго утгаас хамаарах гүйцэтгэлийг анхдагч 0.5 болон Youden index-ээр тохируулсан босго утгаар тооцоолов. Youden босгыг сонгосон 671 зураг дээр нь үнэлсэн тул үр дүн өөдрөгөөр үнэлэгдсэн байж болно. Platt калибрацийг 12 ангилалд өвчтөнөөр бүлэглэсэн таван хэсгийн солбицсон үнэлгээгээр хийж, өвчтөн бүрийг өөрийнх нь өгөгдлөөр тохируулаагүй функцээр үнэлэв. Brier score болон ECE-г тооцов. **Үр дүн:** Загвар сонгоход ашигласан дотоод баталгаажуулалтад 14 ангиллын дундаж AUC 0.8169 байсан бол Монгол өгөгдөл дээр макро AUC 0.7517, микро AUC 0.7762 байв. Дэмжих хэрэгсэл, эмгэг илрээгүй, нэвчдэс, гялтангийн хөндийд шингэн хурах зэрэгт ялган таних чадвар өндөр байсан бол хугарал, уушгины шалчийлт, гялтангийн бусад эмгэгт харьцангуй бага байв. Анхдагч 0.5 босго утгаар макро нарийвчлал 0.8139 боловч макро мэдрэг чанар 0.2347 байсан нь эерэг өөрчлөлтийг орхих эрсдэлтэйг харуулав. Босго утгыг тохируулсны дараа F1 оноо нэвчдэс сүүдэрт 0.006-аас 0.794, зүрх-голтын өргөсөлд 0.019-өөс 0.574 болж нэмэгдэв. Platt калибраци нь зарим өөрчлөлтийн магадлалын тохирлыг сайжруулсан. **Дүгнэлт:** CheXpert дээр сургасан DenseNet121 загвар нь Монгол хүний цээжний рентген зурагт дунд зэрэг шилжин ажиллах чадвартай байсан ч зарим ангилалд гүйцэтгэл буурсан байв. Үүний шалтгааныг энэ судалгаагаар тогтоогоогүй. Тогтмол 0.5 босго утга Монгол өгөгдөлд тохирохгүй байсан бол босго утгыг өөрийн өгөгдөл дээр тохируулснаар F1 оноо эрс сайжирсан нь загварыг орон нутгийн хэрэглээнд тохируулах бодит боломжийг харууллаа.

**Үндэслэл:** Монгол Улсын өргөн уудам нутаг дэвсгэр, хүн амын тархмал суурьшил нь орон нутагт оношилгооны тусламжийг цаг тухайд нь хүргэхэд хүндрэл учруулдаг. Цээжний рентген шинжилгээ нь сүрьеэ, хатгалгаа зэрэг өвчний үед үүсэх өөрчлөлтийг илрүүлж, цаашдын оношилгоог чиглүүлэхэд чухал үүрэгтэй. Иймээс зургийн үнэлгээг дэмжиж, эмгэг сэжигтэй тохиолдлыг эмчийн хяналтад эрэмбэлэн ангилах хиймэл оюуны шийдлийг судлах хэрэгцээ байна.

Гэвч гадаадын өгөгдөлд сургасан загварын гүйцэтгэл өөр хүн ам, эмнэлзүйн орчинд өөрчлөгдөж болох тул Монгол орны нөхцөлд шууд хэрэглэхээс өмнө дотоодын өгөгдөлд үнэлэх нь зүйтэй болохыг Zech JR нарын (2018) судалгаа харуулж байна.[1] Мөн Irvin J нарын (2019) судалгаанд CheXpert өгөгдөлд сургасан загварын эмгэг өөрчлөлтийг ялган таних чадварыг радиологич эмч нарын үнэлгээтэй харьцуулж, тооцсон магадлалын калибрацийг үнэлсэн нь загварын боломж, хязгаарлалтыг олон талаас нь судлахын ач холбогдлыг харуулж байна.[2] Үүнд үндэслэн уг загварын ялган таних чадвар, шийдвэрийн босго болон тооцсон магадлалын найдвартай байдлыг Монгол хүний цээжний рентген зураг үнэлэх нь тухайн хүн амд хэрэглэх боломж, хязгаарлалтыг тодорхойлж, цаашдын эмнэлзүйн туршилтад суурь мэдээлэл бүрдүүлэх ач холбогдолтой.

**Зорилго:** Олон нийтэд нээлттэй CheXpert өгөгдөл дээр DenseNet121 гүн сургалтын загвар сургаж, түүний гүйцэтгэлийг Монгол хүний цээжний рентген зураг дээр үнэлэн, орон нутгийн өгөгдөлд тохируулах боломжийг судлах.

**Арга, аргачлал:** Энэхүү ретроспектив судалгаанд CheXpert өгөгдөлд сургасан гүн сургалтын загварын гүйцэтгэлийг Монгол хүний цээжний рентген зураг үнэлэв. Гадаад үнэлгээнд Улсын Нэгдүгээр Төв Эмнэлэгт 2026 оны дөрөвдүгээр сараас тавдугаар сар хүртэл цээжний рентген шинжилгээнд хамрагдсан 426 өвчтөний дүрс оношилгооны эмчийн дүгнэлттэй 671 зургийг ашиглав.

Радиологич эмч зураг бүрийн өмнө бичигдсэн эмнэлзүйн дүгнэлтээс CheXpert-ийн 14 ангиллын лавлагаа тэмдэглэгээг гаргав. Дүгнэлтэд тухайн ангиллыг эерэгээр тэмдэглэсэн бол 1, сөрөг эсвэл тэмдэглээгүй бол 0 утга оноов. Нэг зураг хэд хэдэн ангилал зэрэг оноох боломжтой байв. Тэмдэглэгээг загварын үр дүнгээс хараат бусаар хийсэн бөгөөд зургийг дахин уншиж үнэлээгүй.

Үнэлгээний нэгж нь зураг байв. Загварын ангилал тус бүрд тооцсон магадлалыг (0-1) лавлагаа тэмдэглэгээтэй харьцуулж, ялган таних чадварыг ROC муруйн доорх талбайгаар (AUC) үнэлэв. Ангилал тус бүрийн AUC болон 14 ангиллын гүйцэтгэлийг нэгтгэсэн макро, микро AUC-ийг тооцов. Үнэлгээнд ангилал бүрд 671, нийт 9 394 (671 × 14) хос утгыг ашиглав. Нэг өвчтөний зургуудын хоорондын хамаарлыг харгалзан AUC-ийн итгэх интервалын тооцоо болон калибрацийн шинжилгээнд өгөгдлийг өвчтөнөөр бүлэглэв.

*Загварын сургалт ба сонголт:* ImageNet өгөгдөлд урьдчилан сургасан жинг эхлэл болгон DenseNet121 архитектурт суурилсан загварыг CheXpert өгөгдөлд сургав.[3] Загвар нь 320 × 320 пикселийн RGB зургийг оролт болгон авч, сигмоид функцээр 14 ангилал тус бүрийн магадлалыг тооцоолов. Эдгээрт эмгэг өөрчлөлтийн ангиллуудаас гадна «эмгэг илрээгүй» болон «дэмжих хэрэгсэл» гэсэн ангилал багтсан.

CheXpert өгөгдлийг өвчтөний түвшинд 80:20 харьцаагаар хувааж, 179 132 зургийг сургалтад, 44 282 зургийг дотоод баталгаажуулалтад ашиглав. Нэг өвчтөний бүх зургийг зөвхөн нэг багцад хуваарилав. Багцын хэмжээг 16 зураг, суралцах хурдыг 0.0001, санамсаргүй тооны үүсгүүрийн эхлэл утгыг 42 гэж тогтоож, сургалтыг таван мөчлөг (epoch) явуулав.

Загвар сонгох шалгуур болгон дотоод баталгаажуулалтын багц дахь ателектаз, зүрх томрол, нэвчдэс сүүдэр, хаван, гялтангийн шингэн гэсэн таван ангиллын AUC-ийн дунджийг ашиглав. Уг үзүүлэлт хамгийн өндөр байсан дөрөвдүгээр мөчлөгийн загварыг сонгов. Сонгосон загварыг Монголын өгөгдөлд дахин сургахгүйгээр гадаад үнэлгээнд ашиглав.

Ялган таних чадвар ба шийдвэрийн босгын үнэлгээ: Загварын ялган таних чадварыг ангилал тус бүрийн ROC-AUC болон нийт ангиллыг нэгтгэсэн макро, микро AUC-аар үнэлэв. Макро AUC-ийг ангиллуудын AUC-ийн ижил жинтэй дунджаар, микро AUC-ийг бүх зураг-ангиллын таамаглалыг нэгтгэн тооцов. Нэг өвчтөний олон зураг хоорондоо хамааралтай байж болохыг харгалзан, AUC-ийн 95%-ийн итгэх интервалыг өвчтөнөөр бүлэглэсэн 1,000 удаагийн bootstrap дахин түүвэрлэлтээр тодорхойлов.

Мэдрэг чанар, өвөрмөц чанар, зөв ангилалтын хувь болон F1 оноог эхлээд анхдагч 0.5 босгоор тооцов. Perkins NJ нарын (2006) судалгаанд авч үзсэн Youden индексийн шалгуурыг ашиглан, ангилал тус бүрд мэдрэг ба өвөрмөц чанарын нийлбэрийг хамгийн их байлгах босгыг сонгож, үзүүлэлтүүдийн өөрчлөлтийг харьцуулав.[4] Босгыг сонгох болон тохируулсны дараах гүйцэтгэлийг үнэлэхэд ижил 671 зургийг ашигласан тул энэ хэсгийг босго тохируулгын боломжийг судалсан шинжилгээ гэж үзэв.

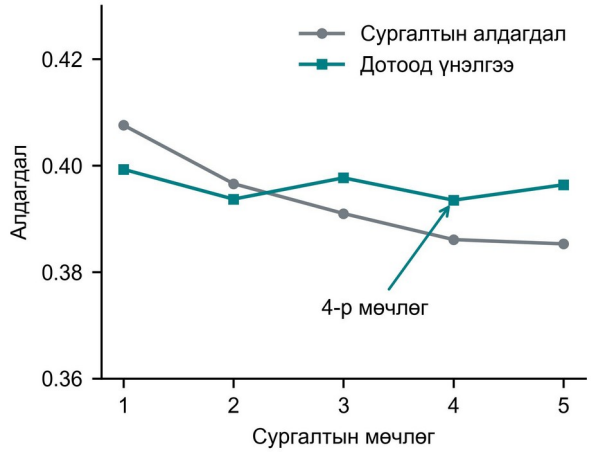
*Магадлалын калибраци:* Загварын тооцсон магадлал бодит эрэг тохиолдлын давтамжтай хэр нийцэж байгааг үнэлж, Guo C нарын (2017) судалгаанд тайлбарласан Platt scaling аргаар тохируулав.[5] Үүнийг өвчтөний түвшинд таван бүлэгт хуваасан хөндлөн баталгаажуулалтаар гүйцэтгэсэн. Тухайлбал, дөрвөн бүлгийн өгөгдлөөр тохируулгын функцийг тогтоож, үлдсэн бүлэгт хэрэглэн, энэ үйлдлийг бүх бүлэгт ээлжлэн давтав. Ийм байдлаар тухайн өвчтөний өгөгдөл өөрийнх нь магадлалыг тохируулах функцийн тооцоонд ороогүй. Эерэг болон сөрөг тохиолдол тус бүр 10-аас доошгүй байсан 12 ангилалд тохируулгын өмнөх ба дараах Brier оноо, ECE үзүүлэлтийн дунджийг харьцуулав.

*Статистик боловсруулалт:* Өгөгдлийг Python орчинд PyTorch, scikit-learn, NumPy, pandas сангуудыг ашиглан боловсруулав. Өвчтөн, зургийн нийт тоо болон ангилал тус бүрийн эерэг зургийн тоо (n), эзлэх хувийг (%) тодорхойлогч статистикаар нэгтгэн үзүүлэв. Загварын гүйцэтгэлийн үзүүлэлтүүдийг цэгэн

үнэлгээгээр, ангилал тус бүрийн AUC-ийг 95%-ийн итгэх интервалын хамт тайлагнав. Статистик таамаглал шалгах тест хийгээгүй тул p утга тооцоолоогүй.

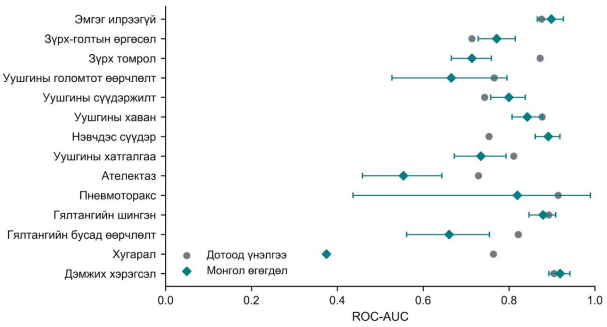
**Үр дүн:** Гадаад үнэлгээнд 426 өвчтөний цээжний рентген шинжилгээний 671 зургийг хамруулав.

*Загварын сонголт:* Дотоод баталгаажуулалтын багц дахь загвар сонгоход ашигласан таван ангиллын дундаж AUC дөрөвдүгээр мөчлөгт хамгийн өндөр байсан тул уг мөчлөгийн загварыг гадаад үнэлгээнд ашиглав. Сургалтын болон дотоод баталгаажуулалтын алдагдлын муруйг Зураг 1-д үзүүлэв.



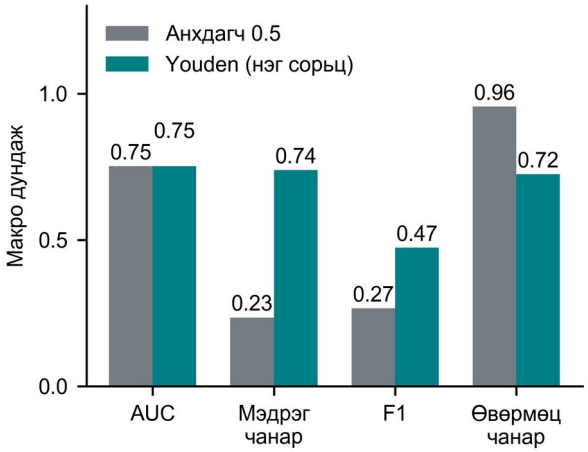

**Зураг 1. Сургалтын болон дотоод үнэлгээний алдаа.** Дөрөвдүгээр мөчлөгийг таван ангиллын дундаж AUC-аар сонгосон.

*Загварын ялган таних чадвар:* Сонгосон загварын дотоод баталгаажуулалтын 14 ангиллын макро AUC 0.8169 байсан бол Монголын өгөгдөлд макро AUC 0.7517, микро AUC 0.7762 байв. Нэвчдэс сүүдрийн AUC 0.8914 (95%-ийн итгэх интервал: 0.8610-0.9187), гялтангийн шингэнийх 0.8793, уушгины хавангийнх 0.8425 байлаа. Эмгэг өөрчлөлтөөс бусад ангиллын хувьд «дэмжих хэрэгсэл»-ийн AUC 0.9190, «эмгэг илрээгүй» ангиллынх 0.8986 байв (Зураг 2). Пневмотораксын ангилалд зургаан эерэг зураг байсан бөгөөд AUC 0.8193, 95%-ийн итгэх интервал 0.4368-0.9894 байв. Хугарлын ангилалд нэг эерэг зураг байсан тул AUC-ийн итгэх интервалыг тайлагнаагүй.



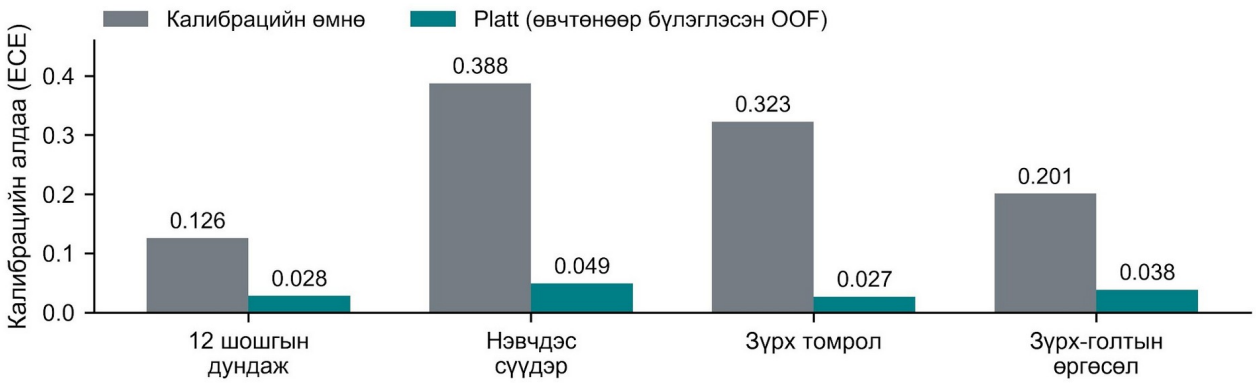

**Зураг 2. Дотоод болон Монголын өгөгдлийн AUC.** Шугамаар Монголын өгөгдлийн өвчтөнөөр бүлэглэсэн 95%-ийн итгэх интервалыг үзүүлэв. Дотоод багцыг загвар сонгоход ашигласан. Хугарлын интервал дүрслээгүй (нэг эерэг зураг).

*Шийдвэрийн босго тохируулгын үр дүн:* Анхдагч 0.5 босгоор 14 ангиллын зөв ангилалтын хувийн макро дундаж 81.39% байв. Нэвчдэс сүүдэртэй 330 зургийн нэгийг зөв илрүүлж, 329-ийг сөрөг гэж ангилсан бол нэвчдэс сүүдэргүй 341 зургийг бүгдийг зөв ангилсан. Ангилал тус бүрд Youden босгыг сонгон, ижил 671 зурагт үнэлэхэд макро мэдрэг чанар 23.47%-аас 73.86%, макро F1 оноо 0.2657-оос 0.4732 болж нэмэгдсэн. Энэ үед макро өвөрмөц чанар 95.60%-аас 72.41% болж буурав (Зураг 3). Нэвчдэс сүүдрийн F1 оноо 0.006-аас 0.794 болсон. Зөвхөн шийдвэрийн босгыг өөрчилсөн тул AUC өөрчлөгдөөгүй.




**Зураг 3. Босго тохируулгын үр дүн: 14 ангиллын дундаж. Youden босгыг сонгосон зураг дээр нь үнэлсэн.**

*Магадлалын тохируулгын үр дүн:* Өвчтөнөөр бүлэглэсэн таван хэсэгт хөндлөн баталгаажуулалтаар үнэлэхэд Platt тохируулгын дараа 12 ангиллын дундаж Brier оноо 0.1614-өөс 0.1207, дундаж ECE 0.1259-өөс 0.0280 болж буурав (Зураг 4). Гэхдээ ECE бүх ангилалд буураагүй бөгөөд гялтангийн шингэний ангилалд нэмэгдсэн.




**Зураг 4. Platt тохируулгын өмнөх ба дараах ECE: 12 ангиллын дундаж, гурван жишээ. Өвчтөнөөр бүлэглэсэн таван хэсгийн үнэлгээ.**

**Хэлцэмж:** Бидний судалгааны үр дүнд DenseNet121 загварын дотоод өгөгдөлд дундаж AUC 0.8169 байсан бол гадаад өгөгдөлд макро AUC 0.7517 болж буурав. Гадаад өгөгдөлд анхдагч 0.5 босгоор макро мэдрэг чанар 0.2347, F1 оноо 0.2657 байсан бол ангилал тус бүрийн босгыг Youden индексээр тохируулсны дараа тус тус 0.7386, 0.4732 болж нэмэгдэв. Энэ үед өвөрмөц чанар 0.9560-аас 0.7241 болсон. Мөн Platt аргаар магадлалын тохируулга хийхэд үнэлсэн 12 ангиллын дундаж Brier оноо 0.1614-өөс 0.1207, ECE 0.1259-өөс 0.0280 болж буурч, тооцсон магадлалын нийцэл сайжрав. Ийнхүү орон нутгийн өгөгдөлд босго болон магадлалыг тохируулах нь эмгэг өөрчлөлтийг илрүүлэх мэдрэг чанар, магадлалын нийцлийг сайжруулах боломжтойг бидний үр дүн харуулж байна.

Шийдвэрийн босгыг тохируулснаар мэдрэг чанар нэмэгдсэн бидний үр дүн Shin нарын (2024) судалгаатай нийцэж байна. Тэд хүүхдийн цээжний рентген зурагт босгыг тохируулсны дараа нэвчдэс сүүдэр илрүүлэх мэдрэг чанар 82.1%-аас 83.4% болж нэмэгдсэнийг мэдээлсэн.[6] Мөн Rudolph нар (2025) амбулаторийн өвчтөнүүдийн зурагт гялтангийн шингэн илрүүлэх босгыг өөрчлөхөд мэдрэг чанар 46.8%-аас 87.2% болсон байна.[7] Эдгээр үр дүн нь босгыг тухайн өгөгдөлд тохируулах замаар эмгэг өөрчлөлтийг илрүүлэх мэдрэг чанарыг нэмэгдүүлэх боломжтой гэсэн бидний дүгнэлтийг дэмжиж байна. Rudolph нарын судалгаанд өвөрмөц чанар мөн буурсан нь манай судалгаанд уг үзүүлэлт 95.6%-аас 72.4% болсон өөрчлөлттэй ижил чиглэлтэй байв.[7] Иймээс босгыг сонгохдоо мэдрэг чанар болон худал эерэг дүгнэлтийн тэнцвэрийг харгалзах нь зүйтэйг харуулж байна.

Магадлалын тохируулгын хувьд бидний олдвор цээжний рентген зурагт хийгдсэн бусад судалгааны үр дүнтэй мөн нийцэж байна. Frenkel, Goldberger нар (2022) ChestX-ray14 өгөгдөлд DenseNet121 загварын гаралтыг confidence-based weight scaling аргаар тохируулахад ECE 4.18%-аас 2.75% болж буурсныг тогтоосон.[8] Харин Salam нар (2026) сүрьеэ илрүүлэх DenseNet121 загварт temperature scaling хэрэглэснээр тест багцын ECE 0.094-өөс 0.052, Brier оноо 0.106-аас 0.097 болж буурсныг мэдээлжээ.[9] Хэрэглэсэн арга нь өөр боловч тохируулгын дараа магадлалын алдаа буурсан эдгээр үр дүн нь манай Platt тохируулгаар магадлалын нийцэл сайжирсан үр дүнтэй ижил хандлагатай байна.

Үүнтэй төстэй үр дүнг бусад дүрс ангилалтын судалгаанд ч мэдээлсэн. Carneiro нар (2020) бүдүүн гэдэсний ургацгийн дурангийн зургийг ангилах загварт temperature scaling хэрэглэснээр калибрацийн алдаа буурсныг тогтоосон.[10] Guo нарын (2017) эмнэлзүйн бус CIFAR-100 өгөгдөлд хийсэн туршилтад мөн энэ аргаар DenseNet40 загварын ECE 10.37%-аас 1.18% болж буурчээ.[5] Эдгээр нь эмнэлзүйн шууд харьцуулалт биш боловч сургалтын дараах тохируулгаар тооцсон магадлалын нийцлийг сайжруулж болохыг харуулснаараа бидний үр дүнг арга зүйн хувьд дэмжиж байна.

Ийнхүү дээрх судалгаануудын олдвор нь босго тохируулгаар мэдрэг чанарыг нэмэгдүүлэх, магадлалын тохируулгаар калибрацийн алдааг бууруулах гэсэн бидний хоёр гол үр дүнтэй чиглэлээрээ нийцэж байна. Энэ нь гадаадын өгөгдлөөр сургасан загварыг орон нутгийн нөхцөлд үнэлэхдээ шийдвэрийн босго болон магадлалын тохируулгыг хамтад нь авч үзэх үндэслэл болж байна.

**Дүгнэлт:** DenseNet121 загвар Монгол хүний цээжний рентген зурагт зарим эмгэг өөрчлөлтийг ялган таних чадвартай байсан бөгөөд ангилал тус бүрийн босгыг тохируулснаар мэдрэг чанар, магадлалын тохируулгаар тооцсон магадлалын нийцэл сайжрав. Эдгээр үр дүнг шинэ өвчтөнүүдийн өгөгдөлд баталгаажуулах нь эмнэлзүйн хэрэглээнд шилжих дараагийн алхам юм.

Энэхүү судалгаа нь эмгэг сэжигтэй зургийг эрэмбэлж, эмчийн үнэлгээг дэмжих систем хөгжүүлэх суурь мэдээлэл бүрдүүлж байна. Цаашдын эмнэлзүйн туршилтаар оношилгоог шуурхай болгох, эмчийн ачааллыг бууруулах, алслагдсан бүсийн оношилгооны хүртээмжийг нэмэгдүүлэхэд үзүүлэх бодит үр нөлөөг тодорхойлох боломжтой.

**Ном зүй:**


1. *Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15(11):e1002683. doi:10.1371/journal.pmed.1002683*
2. *Irvin J, Rajpurkar P, Ko M, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proc AAAI Conf Artif Intell. 2019;33(1):590-597. doi:10.1609/aaai.v33i01.3301590*
3. *Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

2017:4700-4708.
4. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol. 2006;163(7):670-675. doi:10.1093/aje/kwj063
5. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Proc Mach Learn Res. 2017;70:1321-1330.
6. Shin HJ, Han K, Son NH, et al. Optimizing adult-oriented artificial intelligence for pediatric chest radiographs by adjusting operating points. Sci Rep. 2024;14:31329. doi:10.1038/s41598-024-82775-z
7. Rudolph J, Huemmer C, Preuhs A, et al. Threshold optimization in AI chest radiography analysis: integrating real-world data and clinical subgroups. Eur Radiol Exp. 2025;9:95. doi:10.1186/s41747-025-00632-8
8. Frenkel L, Goldberger J. Calibration of medical imaging classification systems with weight scaling. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2022. Cham: Springer; 2022:642-651. (Lecture Notes in Computer Science; vol. 13438). doi:10.1007/978-3-031-16452-1_61
9. Salam A, Andono PN, Purwanto, et al. A reliability-aware analytics framework for multi-vendor Tuberculosis screening from chest radiographs. Healthcare Analytics. 2026;10:100473. doi:10.1016/j.health.2026.100473
10. Carneiro G, Zorron Cheng Tao Pu L, Singh R, Burt A. Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. Med Image Anal. 2020;62:101653. doi:10.1016/j.media.2020.101653

# Evaluation of DenseNet121 decision thresholds and probability calibration on chest radiographs from Mongolian individuals

*Bayasgalant D[1], Orgil-Erdene E[2], Oyunbold L[3], Delgerdalai Kh[4]*

*[1]Department of Radiology, School of Medicine, MNUMS*
*[2]AI Department, AND Global LLC*
*[3]Department of Radiology, First Central Hospital of Mongolia*
*[4]Department of Radiology, MNUMS*
*E-mail: bayasaadc@gmail.com, Tel: +976-80909363*

**Abstract:**
**Background:** Mongolia's vast territory and dispersed population pose challenges to timely diagnostic care. Artificial intelligence may support chest radiograph interpretation and prioritization of suspicious findings for clinician review. Models trained on foreign datasets should therefore be evaluated on Mongolian data before clinical use.
**Aim:** To externally validate a DenseNet121 chest radiograph AI model trained on the public CheXpert dataset using Mongolian clinical data, and to evaluate local adaptation strategies for clinical translation.
**Materials and Methods:** A DenseNet121-based multi-label classifier was trained on CheXpert with ImageNet initialization, 320×320 RGB inputs, and 14 sigmoid outputs, including No Finding and Support Devices. The best checkpoint was selected by internal CheXpert validation and evaluated without fine-tuning on 671 Mongolian chest radiographs whose 14 labels were manually extracted by a radiologist from previous clinical reports. Discrimination was assessed using ROC-AUC, macro AUC, and micro AUC. Threshold-dependent performance was evaluated at the default 0.5 threshold and after label-specific Youden index optimization. Youden thresholds were selected and evaluated on the same 671 images, so improvements may be optimistic. Platt calibration used five-fold patient-grouped cross-validation for 12 eligible labels; each patient was evaluated using a calibrator fitted without their data. Brier score and expected calibration error (ECE) were assessed.
**Results:** The mean 14-label AUC in the internal CheXpert validation set used for model selection was 0.8169. On the Mongolian external validation set, macro AUC was 0.7517 and micro AUC was 0.7762. Strong discrimination was observed for Support Devices (0.9190), No Finding (0.8986), Consolidation (0.8914), Pleural Effusion (0.8793), Edema (0.8425), Pneumothorax (0.8193), and Lung Opacity (0.7997). AUC decreased for Fracture, Atelectasis, Pleural Other, and Cardiomegaly, with the causes of these differences unestablished. At the default 0.5 threshold, macro sensitivity was low (0.2347) despite high macro accuracy (0.8139), indicating that direct transfer may miss positive findings. Local threshold tuning improved F1 scores for Consolidation from 0.006 to 0.794, Edema from 0.509 to 0.600, Lung Opacity from 0.590 to 0.662, and Pleural Effusion from 0.737 to 0.757. Platt scaling improved probability reliability for several labels without systematically increasing AUC.
**Conclusion:** A CheXpert-trained DenseNet121 model showed moderate transferability to Mongolian chest radiographs, with strong performance for several clinically important findings and lower performance for other categories. Threshold optimization and calibration are feasible first-stage adaptation strategies. Larger multi-site Mongolian datasets, consensus labels, independent calibration/test sets, and prospective validation are required before clinical deployment.

**Keywords:** sensitivity; specificity; decision threshold; calibration; external validation