AI analysis used a convolutional neural network with data divided in an 80:20 ratio (development/internal validation), with external validation undertaken using data from the UK Biobank. What is development, internal and external validation?

**Програма ChatGPT сказала:**

Great question! Let's break it down clearly:

**1. Development**

This refers to the **initial training and tuning** phase of the model. During development:

* The model learns patterns from the data (typically 80% of the dataset).
* This part of the data is sometimes further split into:
  + **Training set**: to train the model.
  + **Validation set** (optional within this phase): to fine-tune parameters or prevent overfitting.

So in your case, **80% of the dataset** was used for development (training the convolutional neural network, or CNN).

**2. Internal Validation**

This means testing the model's performance on data **not used for training**, but still from the **same source or dataset**.

* In your case, the remaining **20%** of the original dataset was used for this.
* The goal is to see how well the model performs on "unseen" but **similar** data.

It helps to catch overfitting or errors that wouldn’t show up during training.

**3. External Validation**

This is the **gold standard** for testing a model's generalizability.

* The model is tested on **completely independent data** from a **different source**.
* In your case, this was **data from the UK Biobank**, a large external database.

External validation shows whether the model performs well in real-world or broader applications—not just in the environment it was trained in.

**Summary (in simple terms):**

| **Type** | **Data Source** | **Purpose** |
| --- | --- | --- |
| Development | 80% of the dataset | To train and build the model |
| Internal Validation | 20% of the dataset (same source) | To test on similar but unseen data |
| External Validation | Completely new dataset (UK Biobank) | To test on different, real-world data |

Data were randomly split into sets with a 64:16:20 ratio (train/ test/holdout), with all ECGs belonging to a single patient grouped together to avoid data contamination. Що це?

**Програма ChatGPT сказала:**

Гарне питання! Давай розберемося, що це означає простими словами — і українською, і при потребі з англійською термінологією.

**📊 Розподіл даних (64:16:20 — train/test/holdout)**

Це означає, що весь набір даних був випадково розділений наступним чином:

* **64% — Train (навчальна вибірка)**  
  👉 Тут модель **вчиться**: знаходить закономірності в ЕКГ (ECG) даних.
* **16% — Test (тестова вибірка)**  
  👉 Ця частина використовується для **перевірки моделі під час розробки** — щоб побачити, як вона працює на даних, яких не бачила під час навчання.
* **20% — Holdout (вибірка для внутрішньої валідації)**  
  👉 **Взагалі не використовується під час тренування чи налаштування.**  
  Вона застосовується **тільки один раз — наприкінці**, щоб отримати чесну оцінку того, наскільки модель добре працює **на "нових" даних із того самого джерела**.

**🧬 "ECGs belonging to a single patient grouped together"**

**"ECG однієї людини були згруповані разом"**

Це важливо для **уникнення data contamination (забруднення даних)**.

👉 Уяви, що дані однієї й тієї ж людини потрапили б і в train, і в test.  
Тоді модель ніби **вчиться розпізнавати людину**, а не закономірності в серцевих сигналах.

**Тому всі записи одного пацієнта поміщали в одну групу** (train/test/holdout), щоб уникнути цього ефекту.

**📌 Підсумок простими словами:**

1. Дані поділили на три частини:
   * **64% — навчання**
   * **16% — тестування**
   * **20% — фінальна оцінка (holdout)**
2. Щоб модель не "шпигувала", **всі дані від одного пацієнта тримали разом**, щоб уникнути обману (data leakage).

Хочеш, я намалюю тобі схему цієї структури даних або поясню прикладом, скажімо, з фотками або діагнозами?