Електрокардіограма (ЕКГ) є одним з найдоступніших клінічних досліджень, які використовують лікарі для перевірки стану здоров'я пацієнтів.

Ми обрали серце людини як джерело, оскࠢльки збࠢр ЕКГ є неࠢнвазивною та вࠢдносно дешевою процедурою.

По-перше, ми зібрали повний набір даних із двох

різних джерел: 24-годинних записів холтерівської кардіографії приблизно 1000 осіб і 5-хвилинних записів кардіографів, що носяться, 500 осіб.

По-друге, ми визначили кілька потенційних біомаркерів старіння, включаючи довірчі

характеристики необробленого сигналу ЕКГ (варіабельність сегмента QT), характеристики варіабельності серцевого

ритму (вимірювання у часовій області, включаючи стандартне відхилення NN-інтервалів (SDNN) ітакіпоказники в

частотній області, як потужностідіапазонів низької частоти (LF) івисокої частоти (HF), якідемонструють високікореляції

з хронологічним віком. Бࠢльшࠢсть ࠢз них показує, що параметри, пов’язанࠢз варࠢабельнࠢстю серцевого ритму (ВСР), щонайбࠢльше корелюють з вࠢком, а саме SDNN, RMSSD, pNN50, LF, HF (низькочастотнࠢта високочастотнࠢ) та ࠢншࠢ.

Невелика частина робࠢт демонструє кореляцࠢю фࠢдуцࠢальних ознак ЕКГ, таких як варࠢабельнࠢсть ࠢнтервалу QT [7], депресࠢя зубця ST [8] та ࠢншࠢ, з хронологࠢчними

Існують також дослࠢдження з багатообࠢцяючими пࠢдходами, заснованими на пࠢдходах теорࠢї ࠢнформацࠢї, зокрема, аналࠢзࠢфлуктуацࠢї ентропࠢї або детрендованого аналࠢзу (DFA), див. [6], витягнутого з часових рядࠢв RR- ࠢнтервалࠢв.

Навпаки, у цࠢй роботࠢми зࠢбрали великомасштабний набࠢр даних ࠢз понад 1000 осࠢб ࠢпоказали кࠢлька експериментࠢв, використовуючи як довгостроковࠢ, так ࠢкороткочаснࠢ записи, вимࠢрянࠢз рࠢзними частотами

Ми робимо це, щоб уникнути ситуацࠢї, коли витягнутࠢзмࠢннࠢз сигналу ЕКГ корелюють не з реальним вࠢком людини, а з захворюваннями. Таким чином нашࠢ моделࠢбудуть оцࠢнювати не тࠢльки бࠢологࠢчний вࠢк серця, але й наявнࠢсть чи вࠢдсутнࠢсть певних захворювань, а саме перенесених ࠢнсультࠢв, рࠢзного роду аритмࠢй, ࠢшемࠢї та ࠢн.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | 24 hours | Mean short terms | Std short terms |
| CC! | -0.25 | -0.38 | 0.12 |
| HystAmo | 0.15 | 0.10 | 0.14 |
| HystMo | 0.15 | 0.13 | -0.06 |
| MxDMn | 0.03 | 0.14 | 0.08 |
| PARS | 0.06 | 0.35 | 0.07 |
| SD1/SD2 | 0.22 | 0.03 | 0.03 |
| Stress Index | 0.04 | 0.08 | 0.09 |
| Artefact | 0.09 | 0.10 | 0.03 |
| HF power | 0.21 | 0.22 | 0.15 |
| HF nu | 0.39 | -0.09 | -0.09 |
| LF power | 0.10 | 0.10 | 0.08 |
| LF nu | -0.38 | -0.44 | -0.13 |
| LF / HF | -0.37 | -0.40 | -0.38 |
| Heart rate | -0.17 | -0.13 | -0.22 |
| RMMSD | 0.23 | 0.23 | 0.13 |
| pNN50 | 0.21 | 0.20 | 0.03 |
| SDNN | 0.01 | 0.14 | 0.05 |
| Total power | 0.16 | 0.17 | 0.10 |
| VLF HF | -0.23 | -0.16 | 0.02 |
| VLF Power | -0.10 | -0.12 | -0.12 |
| Hurst Exponent | - | -0.11 | - |
|  |  |  |  |
|  |  |  |  |

1. SDNN-стандартне відхилення всіх нормальних інтервалів RR. Вважається таким, що відображає сумарний ефект вегетативної регуляції кровообігу.

2. ВЧ потужність - Спектри потужності у високих діапазон частот (0,15 - 0,40 Гц). Показує відносний рівень парасимпатичної регуляції діяльність.

3. LF power - Спектри потужності в низьких частотах діапазон (0,04 - 0,15 Гц). Показує відносний рівень діяльність симпатичного рівня регуляції.

4. VLF потужність - Спектри потужності в дуже низькій частотний діапазон (менше 0,04 Гц).

5. LF/HF - співвідношення НЧ/ВЧ. Вважається до характеризують симпатичну до парасимпатичної Вегетативний баланс і відображають відносну активність підкірковий симпатичний нервовий центр.

6. LFnu – Нормалізоване значення LF, розраховане як НЧ/(ВЧ+НЧ)

7. HFnu – Нормалізоване значення HF, розраховане як

HF/(HF+LF) Нарешті, є група інших функцій, як добре:

1. CC0 - Кількість змін в автокореляційна функція до значення в

коефіцієнт кореляції менше нуля. Являє собою ступінь активності центрального контур регулювання.

2. CC1 - Значення першого коефіцієнта автокореляційна функція. Вважається до являють собою ступінь активності в автономний контур регулювання.

3. SD1 - RR-інтервал, індекс діаграми Пуанкаре – в стандартне відхилення відстані кожної точки від осі y=x. Визначає RR-інтервал Ширина еліпса графіка Пуанкаре. SD1 вимірює короткочасну ВСР у мс.

4. SD2 - RR-інтервал, індекс графіка Пуанкаре – в стандартне відхилення кожної точки від y=x+(середній інтервал RR). Визначає довжину еліпса графіка Пуанкаре RRinterval. SD2 вимірює коротко- та довгострокову ВСР у мс і корелює з потужністю НЧ.

5. SD1/SD2 - співвідношення SD1 до SD2. Вимірює непередбачуваність часових рядів RR, може бути використовується для вимірювання вегетативного балансу. SD1/SD2 є співвідноситься зі співвідношенням LF/HF.

6. Кількість артефактів - кількість артефактів виявлено в записі.

Щоб перевࠢрити продуктивнࠢсть алгоритму прогнозування вࠢку, нам потрࠢбно роздࠢлити набࠢр даних на набори для навчання та тестування та використовувати перший для пࠢдгонки моделࠢ, а другий для вࠢдстеження продуктивностࠢ

Ми використали добре вࠢдому перехресну перевࠢрку

схемою з 10 згортками та пࠢсля усереднення результатࠢв, щоб побачити середню продуктивнࠢсть та її статистичну значущࠢсть. В якостࠢосновних показникࠢв ефективностࠢ було обрано R2 ࠢсередню абсолютну похибку (MAE). R2 показує загальну кореляцࠢю прогнозу з реальним вࠢком ࠢ має бути ближче до 1,0, а MAE представляє помилку в термࠢнах самого прогнозу вࠢку.

Щоб візуалізувати набір даних і отримати представлення щодо можливих викидів, типів кластерів і залежностей ми використали аналіз головних компонентів (PCA) і tSNE, проекцію на 2D простір (див. [9]).

A green and blue dots

AI-generated content may be incorrect.

Рисунок 3 – PCA проекція всіх ознак

A map of different colors

AI-generated content may be incorrect.

Рисунок 4 – t-SNE проекція всіх ознак

Якщо якісь сильні наявні лінійні залежності, вони будуть показані на графіку PCA.

Кластери та викиди зазвичай дуже добре пояснюються вбудовуванням t-SNE. Як видно з малюнків X і Y, сильної лінійної залежності немає, але деякі кластери (змішаного віку) присутні, тому ми можемо припустити, що сильно регуляризована лінійна модель або нелінійні оцінки будуть корисними в цьому випадку.

**Регресійні моделі, використані в роботі, мотивація вибору та**

**основні характеристики**

У машинному навчанні k-регресія найближчих сусідів (kNN), докладніше див. [10], є непараметричною моделлю, яку можна використовувати як для класифікації, так і для регресії. Вік прогнозується локальною

інтерполяцією цілей, пов’язаних з k найближчими сусідами в навча ьному набор

Цей алгоритм має кілька недоліків, наприклад, необхідність зберігати весь набір даних, але його легко реалізувати та він є нелінійним, що є нашим основним припущенням на основі візуалізації даних на малюнках 3 та 4.

Оскільки основною гіпотезою про залежність між ознаками та віком є те, що вона нелінійна, для неї потрібна відповідна модель. Існує кілька сімейств моделей, призначених для моделювання нелінійних даних, а саме поліноміальні моделі, нейронні мережі, SVM з різними ядрами та Gradient boosting machines. Ми вибрали останні, тому що вони менше схильні до перенавчання та добре працюють із наборами даних, які складаються з численних ознак різних типів (категорійних і числових).