ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

Харківський національний університет радіоелектроніки

каф. ЕОМ

Інформаційна система обробки медичних даних

Ст. групи КІУКІ-21-3
Максименко Є. Р.

Керівник
ас. Романюк О.С.

Мета та задачі роботи

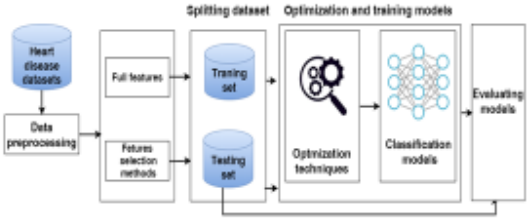
Метою кваліфікаційної роботи є розробка моделі прогнозування серцевих захворювань в якості основи інформаційної системи обробки медичних даних

Задачі:

- Провести аналіз існуючих моделей машинного навчання
- Розробити власну модель для прогнозування серцевих захворювань
- Обрати шкалу оцінювання моделей
- Провести тестування моделей на різних наборах даних

Модель прогнозування серцевих захворювань

- У рамках створення інформаційної системи для обробки даних медичних оцінюємо три підходи: класичний підхід машинного навчання, підхід гібридних моделей та запропоновану модель. Ці моделі застосовуються до повного набору ознак та вибраного набору ознак.
- Запропонована модель прогнозування серцевих захворювань має кілька етапів, включаючи збір даних, попередню обробку даних, розділення даних, вибір ознак та моделі оцінки, як показано на рисунку.



3

Гібридні архітектури моделей

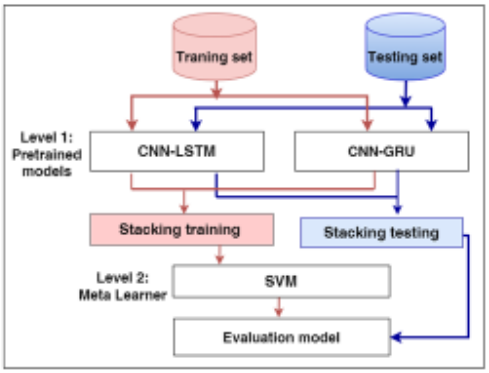
- В ході роботи нами було досліджено та запрапоновано дві гібридні моделі: CNN-LSTM та CNN-GRU для прогнозування серцевих захворювань. Структури гібридних моделей проілюстровано на рисунку.
- Перша модель – це CNN-LSTM, яка поєднує CNN з LSTM та складається зі згорткового шару, шару максимального об'єднання, шару LSTM, шару вирівнювання, повністю зв'язного та вихідного шару;
- Друга модель – CNN-GRU, яка поєднує CNN з GRU. Архітектура складається зі згорткового шару, шару максимального об'єднання, шару GRU, шару вирівнювання, шару повного зв'язку та вихідного шару.

Архітектура гібридних моделей CNN-LSTM та CNN-GRU, що використовуються для прогнозування серцевих захворювань

4

Запропонована модель ансамблю стекування

- В роботі запропонована модель яка має 2 рівні: Рівень 1 та Рівень 2, як показано на рисунку.
- Рівень 1 починається із завантаження попередньо навчених моделей гібридних моделей CNN-LSTM та CNN-GRU, а шари моделей заморожуються, за винятком останніх шарів.
- Моделі передбачають вихідні ймовірності навчального набору та згодом інтегрують їх у стекове навчання. По-друге, моделі оцінюють вихідні ймовірності тестового набору та агрегують їх у стековому тестуванні.
- На рівні 2 SVM, як мета-навчання, навчається та оптимізується за допомогою стекового навчання та пошуку в сітці відповідно, одночасно отримуючи кінцеві результати за допомогою стекового тестування.



5

Оцінювання моделей

- Найчастіше використовуються такі метрики ефективності класифікації, як точність (ACC), прецизійність (PRE), повнота (REC) та F1-оцінка (F1).
- На відміну від істинно позитивного результату (TP), який означає, що людина хвора, а тест позитивний, істинно негативний результат (TN) показує, що людина здорова, а результат негативний. Хибнопозитивні результати – це тести, які виявляються позитивними, навіть коли суб'єкт здоровий (FP). Коли тест негативний, але суб'єкт хворий, це називається хибнонегативним результатом (FN).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN},$$

$$Precision = \frac{TP}{TP + FP}$$

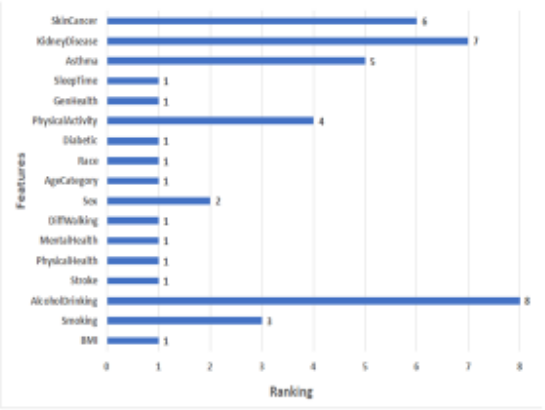
$$Recall = \frac{TP}{TP + FN}$$

$$F1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

6

Результати набору даних 1
Результати вибору ознак.

- В експериментах ми використовували RFE для вилучення важливих ознак з набору даних про серцеві захворювання, присвоївши ранжування кожній ознакі. Критичні ознаки мають ранжування 1, а найменш важливі – 8.
- Ранжування ознак показано на рисунку. Ми бачимо, що 10 найважливіших ознак мають ранг 1: ІМТ, інсульт, фізичне здоров'я, психічне здоров'я, відмінності в ходьбі, вікова категорія, раса, діабетик, генеалогічне здоров'я та час сну. Найменш важлива ознака має рейтинг 8 – вживання алкоголю.



7

Результати набору даних 1
Результати застосування моделей

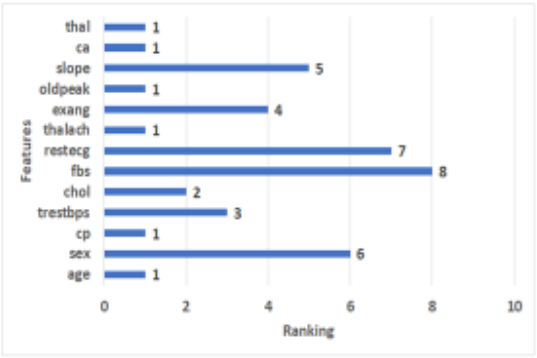
- В роботі представлені ACC, PRE, REC та F1 моделей ML, гібридних моделей, а також запропонована модель для набору даних 1.
- У гібридних моделях CNN-LSTM та CNN-GRU деякі параметри були адаптовані: batch_size 500, epoch = 50, learning rate = 0.00004, а оптимізатором, що використовується, є Adam. Деякі з найкращих значень гіперпараметрів CNN-LSTM та CNN-GRU, які були вибрані KerasTuner.
- В таблиці наведено результати застосування машинного навчання, гібридних моделей та запропонованої моделі з повним набором функцій та вибраними функціями за допомогою радіочастотної епітеліальної функції (RFE) до набору даних 1 щодо захворювань серця.

Підходи	Моделі	Особливості	Матриця продуктивності			
			ACC	PRE	REC	F1
Звичайний підхід до машинного навчання	RF	Повні функції	75.32	75.44	75.32	75.33
		Вибрані функції	73.02	73.06	73.02	73.03
	LR	Повні функції	75.60	75.60	75.60	75.60
		Вибрані функції	73.58	73.60	73.58	73.59
	DT	Повні функції	67.28	67.26	67.28	67.27
		Вибрані функції	65.76	65.76	65.76	65.7
	NB	Повні функції	60.87	64.98	60.87	56.69
		Вибрані функції	60.84	64.97	60.84	56.63
	KNN	Повні функції	73.16	73.47	73.16	73.16
		Вибрані функції	72.59	72.92	72.59	72.59
Гібридні моделі	CNN-LSTM	Повні функції	76.64	76.9	76.64	76.65
		Вибрані функції	75.22	75.42	75.22	75.22
	CNN-GRU	Повні функції	75.63	75.65	75.63	75.58
		Вибрані функції	74.07	74.23	74.07	74.08
Запропонована модель	Stacking SVM	Повні функції	78.81	78.1	78.81	78.81
		Вибрані функції	77.42	77.99	77.42	77.39

8

Результати набору даних Клівленда.
Результати вибору ознак.

- В експериментах ми використовували RFE для вилучення важливих ознак з набору даних Клівленда. Він призначає ознакам значення рангу, де критичні ознаки мають рейтинг 1, а найменш важливі ознаки - рейтинг 8.
- Рейтинг ознак показано на рисунку Ми бачимо, що 8 найважливіших ознак мають рейтинг 1: age, cp, thalach, oldpeak, ca та thal. Найменш важлива ознака має рейтинг 8, що дорівнює fbs.



9

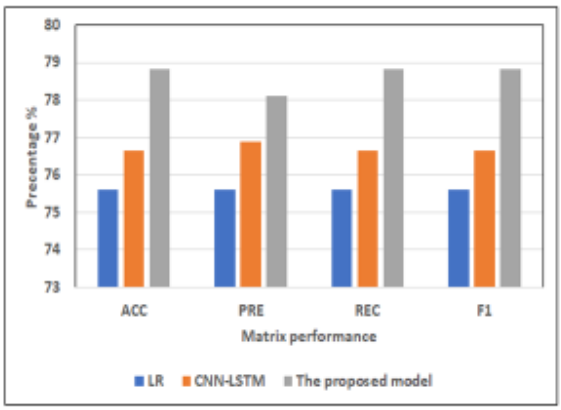
Результати набору даних Клівленда.
Результати застосованих моделей.

Результат застосування моделей з повними та вибраними ознаками для набору даних Клівленда.

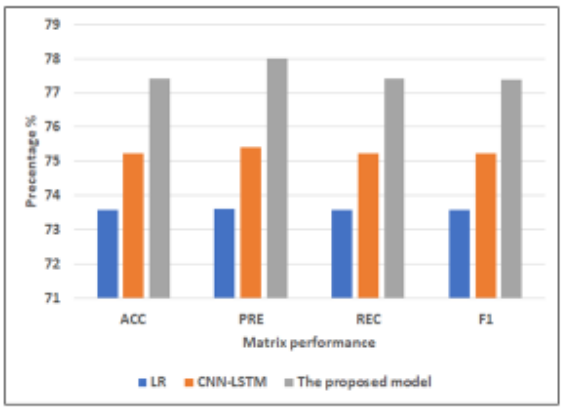
Підходи	Моделі	Особливості	Метрики продуктивності			
			ACC	PRE	REC	F1
Звичайний підхід до машинного навчання	RF	Повні функції	86.34	86.34	86.34	86.34
		Вибрані функції	82.93	82.99	82.93	82.91
	LR	Повні функції	67.52	67.43	67.5	67.18
		Вибрані функції	78.17	78.19	78.17	78.14
	DT	Повні функції	82.44	82.46	82.44	82.44
		Вибрані функції	81.95	82.01	81.95	81.93
	NB	Повні функції	60.00	60.05	60.00	59.74
		Вибрані функції	64.88	64.90	64.88	64.88
	KNN	Повні функції	60.00	60.25	60.00	59.92
		Вибрані функції	66.34	66.62	66.34	66.29
Гібридні моделі	CNN-LSTM	Повні функції	89.76	89.96	89.76	89.75
		Вибрані функції	86.54	86.41	86.34	86.54
	CNN-GRU	Повні функції	88.29	89.06	88.29	88.26
		Вибрані функції	85.85	86.92	85.85	85.78
Запропонована модель	Stacking SVM	Повні функції	97.17	97.42	97.17	97.15
		Вибрані функції	91.22	91.29	91.22	91.22

10

Аналіз результатів. Набір даних 1



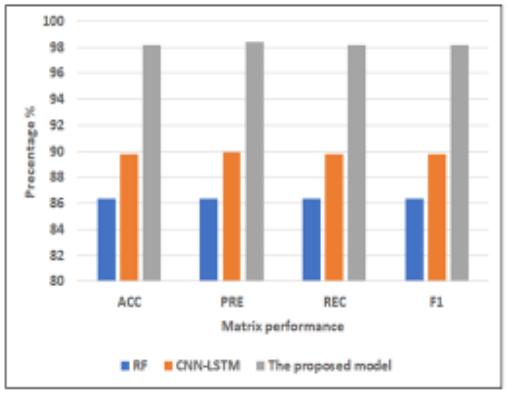
Найкращі моделі для застосування моделей з повним набором функцій для набору даних 1.



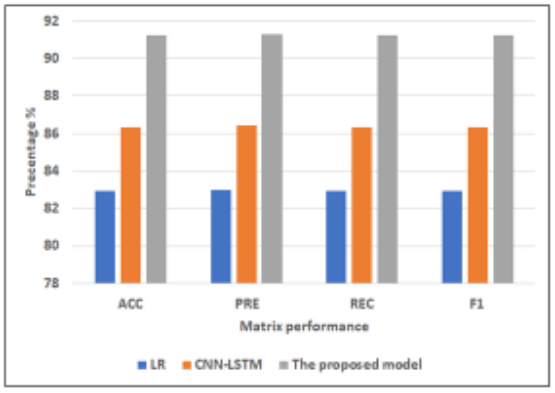
Найкращі моделі для застосування моделей з вибраними ознаками для набору даних 1.

11

Аналіз результатів.Набір даних Клівленда



Найкращі моделі для застосування моделей з повним набором ознаками для набору даних 2.



Найкращі моделі для застосування моделей з вибраними ознаками для набору даних 2.

12

Висновки

В ході кваліфікаційної роботи було розроблено модель прогнозування серцевих захворювань в якості основи інформаційної системи обробки медичних даних

Вирішені наступні задачі:

- Проведено аналіз існуючих моделей машинного навчання
- Розроблено власну модель для прогнозування серцевих захворювань
- Обрано шкалу оцінювання моделей
- Проведено тестування моделей на різних наборах даних

13