Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет комп’ютерної інженерії та управління
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

Рівень вищої освіти перший (бакалаврський)

Інформаційна система обробки медичних даних

(тема)

Виконав:
здобувач 4 року навчання,
групи КІУКІ-21-3
Єгор МАКСИМЕНКО
(власне ім’я, прізвище)

Спеціальність
123 «Комп’ютерна інженерія»
(код і повна назва спеціальності)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма
Комп’ютерна інженерія
(повна назва освітньої програми)

Керівник: ас. Олександр РОМАНЮК
(посада, власне ім’я, прізвище)

Допускається до захисту

Завідувач кафедри ЕОМ _____ Андрій КОВАЛЕНКО
(підпис) (власне ім’я, прізвище)

2025 р.

Харківський національний університет радіоелектроніки

Факультет комп’ютерної інженерії та управління

Кафедра електронних обчислювальних машин

Рівень вищої освіти перший (бакалаврський)

Спеціальність 123 «Комп’ютерна інженерія»
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Комп’ютерна інженерія
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
“____” _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Максименку Єгору Руслановичу
(прізвище, ім’я, по батькові)

1. Тема роботи Інформаційна система обробки медичних даних

затверджена наказом по університету від “ 26 ” травня 2025 р. № 424 Ст

2. Термін подання здобувачем роботи до екзаменаційної комісії 17 червня 2025 р.

3. Вхідні дані до роботи прогнозування, глибоке навчання, CNN-LSTM, CNN-GRU стекінг, медичні дані, інформаційна система Python

4. Перелік питань, що потрібно опрацювати у роботі

Основні положення

Методологія запропонованої інформаційної системи

Тестування інформаційної системи

Висновки

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій 13

6. Консультанти розділів роботи (заповнюється за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / терміни виконання етапів роботи	Примітка
1	Отримання теми кваліфікаційної роботи	26.05	
2	Аналіз літератури	27.05-29.05	
3	Побудова системи	28.05-10.06	
4	Тестування системи та отримання результатів	11.06-12.06	
5	Формування пояснювальної записки	13.06-14.06	
6	Перевірка на плагіат	15.06-17.06	
7	Рецензування роботи	17.06	
8	Подача роботи в ЕК	18.06	
9	Захист роботи	24.06	

Дата видачі завдання " 25 " травня 2025 р.

Здобувач _____
(підпис)

Керівник роботи _____ ас. Олександр РОМАНЮК
(підпис) (посада, власне ім'я, прізвище)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 51 с., 8 рис., 6 табл., 1 дод., 43 джерел.

СЕРЦЕВІ ЗАХВОРЮВАННЯ, ПРОГНОЗУВАННЯ, ГЛИБОКЕ НАВЧАННЯ, CNN-LSTM, CNN-GRU, СТЕКІНГ, МЕДИЧНІ ДАНІ, ІНФОРМАЦІЙНА СИСТЕМА

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

У ході виконання кваліфікаційної роботи представлено ансамблеву модель глибокого стекінгу для підвищення ефективності прогнозування серцевих захворювань. Запропонована інформаційна система базується на інтеграції двох гібридних глибоких моделей — CNN-LSTM та CNN-GRU — із використанням класифікатора SVM як мета-моделі. Для підвищення точності та зменшення надлишковості ознак застосовано метод рекурсивного виключення ознак (RFE). Експериментальні дослідження проводились на двох наборах даних про серцеві захворювання, включаючи відомий набір Cleveland, із порівнянням результатів роботи моделі з класичними алгоритмами машинного навчання (LR, RF, K-NN, DT, NB).

За результатами тестування, запропонована модель показала найвищу точність (ACC), повноту (REC), точність (PRE) та значення F1-міри серед усіх розглянутих моделей. Зокрема, для набору Cleveland досягнуто точності 97,17%, що перевищує результати аналогічних підходів, описаних у науковій літературі.

ABSTRACT

Bachelor’s thesis: 51 pages, 8 figures, 6 tables, 1 appendices, 43 sources.

HEART DISEASE, PREDICTION, DEEP LEARNING, CNN-LSTM, CNN-GRU, STACKING, MEDICAL DATA, INFORMATION SYSTEM.

The major goal of this thesis is to develop a heart disease prediction model as the foundation of a medical data processing information system.

This bachelor's qualification work presents a deep stacking ensemble model designed to improve the effectiveness of heart disease prediction. The proposed information system is based on the integration of two hybrid deep learning models – CNN-LSTM and CNN-GRU – with an SVM classifier serving as the meta-learning model. To enhance accuracy and reduce feature redundancy, the Recursive Feature Elimination (RFE) method was applied. Experimental evaluations were conducted on two heart disease datasets, including the well-known Cleveland dataset, and the proposed model was compared against several classical machine learning algorithms (LR, RF, K-NN, DT, NB).

The results demonstrate that the proposed model achieved the highest performance across all evaluated metrics, including accuracy (ACC), recall (REC), precision (PRE), and F1-score. Specifically, on the Cleveland dataset, it achieved an accuracy of 97.17%, outperforming existing methods described in the literature.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

CNN – Convolutional Neural Network

LSTM – Long Short-Term Memory

GRU – Gated Recurrent Unit

SVM – Support Vector Machine

RFE – Recursive Feature Elimination

F1 –F1-оцінка

ІС – Інформаційна система

ВСТУП

Протягом історії людство вражало багато епідемій, забираючи життя. У наш час було зазначено, що серцеві захворювання є одними з найсмертоносніших захворювань, з якими зіткнулося людство в сучасний період. Поширення шкідливих звичок, таких як куріння, переїдання та відсутність фізичної активності, сприяло зростанню захворюваності на серцеві захворювання. Смертельна особливість серцевих захворювань, яка отримала назву «тихий вбивця», полягає в тому, що вони часто не мають жодних очевидних ознак заздалегідь. Як результат, необхідні дослідження для розробки перспективної моделі ранньої діагностики серцевих захворювань за допомогою простих даних та симптомів. Метою статті є пропонування моделі глибокого стекування ансамблю для підвищення ефективності прогнозування серцевих захворювань. Запропонована модель ансамблю інтегрує дві оптимізовані та попередньо навчені гібридні моделі глибокого навчання з машиною опорних векторів (SVM) як моделлю мета-навчання. Перша гібридна модель - це згорткова нейронна мережа (CNN) - довга короткочасна пам'ять (LSTM) (CNN-LSTM), яка інтегрує CNN та LSTM. Другою гібридною моделлю є CNN-GRU, яка інтегрує CNN з рекурентним блоком (GRU). Для процесу оптимізації вибору ознак також використовується метод рекурсивного виключення ознак (RFE). Запропонована модель була оптимізована та протестована з використанням двох різних наборів даних про захворювання серця. Запропонований ансамбль порівнюється з п'ятьма моделями машинного навчання, включаючи логістичну регресію (LR), випадковий ліс (RF), K-найближчих сусідів (K-NN), дерево рішень (DT), наївний байєсівський метод (NB) та гібридні моделі. Крім того, для оптимізації ML, DL та запропонованих моделей використовуються методи оптимізації. Результати, отримані за допомогою запропонованої моделі, досягли найвищої продуктивності з використанням

повного набору ознак.

Крім високої точності, модель демонструє відмінну стабільність та узагальнюваність результатів на різних вибірках. Це свідчить про її ефективність як інструменту для підтримки прийняття рішень у медичній практиці. Запропонований підхід може бути легко адаптований для інших типів захворювань, що потребують ранньої діагностики на основі обмеженої кількості клінічних даних.

1 ОСНОВНІ ПОЛОЖЕННЯ

1.1 Теоретичні основи дослідження

Серцево-судинні захворювання є одними з найпоширеніших захворювань, які зберігалися в минулому та зросли й поширилися в наш час. Причини зростання їхньої захворюваності різноманітні, особливо в сучасну епоху. Діабет, гіпертонія, холестерин, нерегулярне серцебиття та багато інших клінічних ознак – це деякі біологічні маркери та фактори ризику, необхідні для діагностики серцевих захворювань. Всесвітня організація охорони здоров'я (ВООЗ) стверджує, що однією з основних і найпоширеніших причин смерті в усьому світі є серцево-судинні захворювання, які можуть мати кілька форм, таких як ішемічна, гіпертонічна та судинна хвороба серця [1], і було показано, що серцево-судинні захворювання щороку вбивають 17,9 мільйона пацієнтів. Крім того, нездорова поведінка, яка призводить до надмірної ваги, ожиріння та гіпертонії, підвищує ризик серцевих захворювань [1]. Крім того, серце є одним з найважливіших органів людського організму. Воно в першу чергу відповідає за безперервність перекачування крові, необхідної для роботи решти людського тіла. Однак серцю важко підтримувати таку ж ефективність протягом усього життя людини. Серце схильне до багатьох проблем, які можуть виникати з різних причин, таких як погане здоров'я та харчові звички або старіння [2]. Тому пошук методів і методик, що дозволяють раннє виявлення або навіть прогнозування потенційних проблем із серцем, став неминучим. Це може допомогти лікарям та медичним організаціям зменшити проблеми та ускладнення цього захворювання.

Штучний інтелект (ШІ) на основі машинного навчання (МН) та глибокого навчання (ГН) відіграв ключову роль в оцінці медичних даних для допомоги в діагностиці захворювань та визначенні відповідного лікування.

Він використовується для автоматичного пошуку закономірностей з клінічних даних, а потім для міркування на основі клінічних даних для прогнозування раннього ризику для пацієнтів, таких як серцеві захворювання [3], рак [4,5] та COVID-19 [6,7]. Нещодавно алгоритми глибокого навчання, такі як LSTM, GRU, CNN, та гібридні моделі цих алгоритмів відіграли важливу роль у посиленні та підвищенні рівня прогнозування серцевих захворювань за допомогою різних шарів, які можуть збирати глибші ознаки [8-11]. Нещодавно автори використовували ансамблеве навчання для підвищення продуктивності цих моделей у сфері охорони здоров'я [12]. Ансамблеве навчання поєднує рішення різних базових класифікаторів, використовуючи багато методів, таких як голосування або усереднення, для покращення остаточного рішення [13]. Ансамблеві алгоритми можна розділити на три гілки: бустинг [14], стекінг [15] та бэґґінг [16]. Стекінг ансамблю вважається найкращим методом побудови ансамблевих моделей, оскільки він базується на мета-навчанні, яке навчається з даних, як зважувати базові класифікатори та комбінувати їх найкращим чином для оптимізації продуктивності результуючої моделі. Стекінг ансамблю оптимізує набір гетерогенних базових моделей та комбінує їхні рішення за допомогою мета-навчання [15].

У цьому дослідженні ми запропонували оптимізовану модель ансамблю, яка об'єднала дві попередньо навчені гібридні моделі CNN-LSTM та CNN-GRU з мета-навчальним методом (SVM) для покращення ефективності прогнозування серцевих захворювань. Крім того, для вибору найбільш інформативних ознак з двох наборів даних про серцеві захворювання було використано рекурсивне виключення ознак (RFE). Наш внесок можна підсумувати наступним чином:

- ми запропонували дві гібридні моделі з гетерогенними архітектурами: CNN-LSTM та CNN-GRU, які були запропоновані та оптимізовані.

- ми запропонували модель ансамблю стекування, яка об'єднала

попередні попередньо навчені гібридні моделі CNN-LSTM та CNN-GRU. Найкращий мета-класифікатор для навчання було обрано на основі експериментальних результатів. Алгоритм SVM досяг найкращих результатів як мета-класифікатор для визначення найкращих ваг базових класифікаторів;

- ми порівняли запропоновану модель з різними моделями машинного навчання, використовуючи два еталонні набори даних про серцеві захворювання;

- запропонована модель значно перевершила всі інші моделі та досягла найкращих результатів.

1.2 Огляд літературних джерел

Машинне навчання та глибоке навчання використовувалися для прогнозування серцевих захворювань. Наприклад, в [17] запропонували гібридну модель, яка поєднує DT та RF для прогнозування серцевих захворювань, використовуючи набір даних Клівленда. Вони порівняли ефективність гібридної моделі з ефективністю DT та RF.

В [18] застосували різні алгоритми машинного навчання: SVM, DT, LR, NB, адаптивне підвищення (AdaBoost), стохастичний градієнтний спуск (SGD), RF, машину градієнтного підвищення (GBM) та класифікатор додаткового дерева (ETC), використовуючи набір даних Клівленда про серцеві захворювання для аналізу серцевої недостатності. Результати показали, що ETC показав найкращу продуктивність та перевершив інші моделі. В [19] використовували алгоритми машинного навчання для прогнозування серцевих захворювань: NB, RF, DT, K-NN та SVM. Результати показали, що RF має найкращу ефективність моделі.

Багато авторів застосовували методи вибору ознак з моделями ML та DL для прогнозування захворювань серця. Наприклад, в [20] використовували методи вибору ознак Chi2, ReliefF, симетричної невизначеності (SU) та PCA для вилучення важливих ознак з чотирьох

наборів даних про захворювання серця. Вони застосували BayesNet, Logistic, Stochastic Gradient Descent (SGD) та KNN Adaboost до повних та вибраних ознак. Результат показав, що модель BayesNet була зареєстрована як найкраща з використанням вибору ознак Chi-2 порівняно з іншими моделями.

В [21] використовували алгоритм Lasso для вибору ознак з набору даних про захворювання серця. Вони застосували моделі ML та DL: LR, KNN, SVM, RF, DT та ANN відповідно. Результати показали, що ANN має найкращу продуктивність порівняно з моделями ML. В роботі [22] використовували KNN, MLP, SVM та J48 для виявлення захворювань серця. Набори даних були зібрані з різних джерел. Автори застосували різні стратегії вибору ознак, включаючи класифікатор додаткового дерева, класифікатор з градієнтним підвищенням, випадковий ліс, рекурсивне видалення ознак та класифікатор з підвищенням XG.

У дослідженні [23], щоб підвищити точність прогнозування, автори запропонували гібридну модель голосування на основі NB та LR. Вони використовували k-NN, DT, NB, LR, SVM, нейронну мережу (NN) та гібридну модель для вибору значущих характеристик з набору даних про хвороби серця Клівленда. Гібридна модель отримала найкращу продуктивність порівняно з іншими моделями.

В [24] використовували моделі DT, LR, NB, SVM та RF з методами вилучення ознак з набору даних про хвороби серця Клівленда для прогнозування захворювань серця. Результати показали, що LR та SVM з методами вибору ознак мали кращу точність, ніж інші моделі. В [25] розробили GRU та RF на основі моделей (GRU-RF) для виявлення захворювань серця. GRU-RF порівнювали з алгоритмами RF, GRU, KNN та DNN, і досягли найкращої продуктивності.

В роботі [26] запропонували гібридну модель LSTM–GRU та порівняли її з DT, RF, LR, LSTM та GRU для прогнозування серцевих захворювань. Вони використовували набір даних з лікарні Чонан Університету Сунчунхян

у Кореї для навчання та тестування моделей. Вони покращили моделі продуктивності на основі коригування гіперпараметрів, кількості первинних даних пацієнтів та вхідних параметрів. Результати показують, що порівняно з іншими моделями модель GRU перевершує інші. У дослідженні [27] автори використовували налаштування гіперпараметрів LSTM та GRU для підвищення продуктивності алгоритмів. Результати продемонстрували, що GRU забезпечує кращу точність, ніж LSTM за всіма критеріями.

Автори використовували ансамблеві моделі для прогнозування серцевих захворювань. Наприклад, в [28] застосували LR, SVM, DT, K-NN, GNB та ансамблеві моделі, використовуючи набір даних, зібраний з набору даних про серцеві захворювання UCI. Вони використовували ансамблеві моделі голосування та усереднення, побудовані шляхом об'єднання вищезазначених моделей ML. Результати показали, що ансамблева модель мала найкращі результати порівняно з іншими моделями. В роботі [29] використовували RF, SVM, K-NN, LSTM, модель ансамблю жорсткого голосування та GRU для прогнозування серцевих захворювань. Результати показали, що модель ансамблю жорсткого голосування продемонструвала вищу точність порівняно з іншими моделями.

В досліджені [30] запропонували гібридні моделі, які інтегрували бустинг та бегінг з традиційними моделями машинного навчання: KNN, DT та RF. Гібридні моделі: метод бегінгу K-NN (KNNBM), метод DT-бегінгу (DTBM), AdaBoost (AB) та метод випадкового лісу бегінгу (RFBM) були застосовані до наборів даних про серцеві захворювання. Рельєф, найменше абсолютне скорочення та оператор вибору були трьома підходами до вибору ознак, які вони використовували (LASSO). У порівнянні з іншими моделями, модель RFBM показала найкращу продуктивність.

Попередні дослідження не використовували ансамблеве стекінгове моделювання на основі гетерогенних гібридних моделей глибокого навчання для прогнозування серцевих захворювань. Крім того, більшість попередніх досліджень використовували базу даних серцевих захворювань Клівленда

для проведення цього експерименту. У нашій роботі ми використовували новий великий набір даних про серцеві захворювання та запропонували моделі ансамблевого стекінгу, засновані на оптимізації різних гетерогенних гібридних моделей: CNN-LSTM та GRU-LSTM.

2 МЕТОДОЛОГІЯ ЗАПРОПОНОВАНОЇ ІНФОРМАЦІЙНОЇ СИСТЕМИ

У цій кваліфікаційній роботі 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 2.1. Кожен етап детально описано нижче.

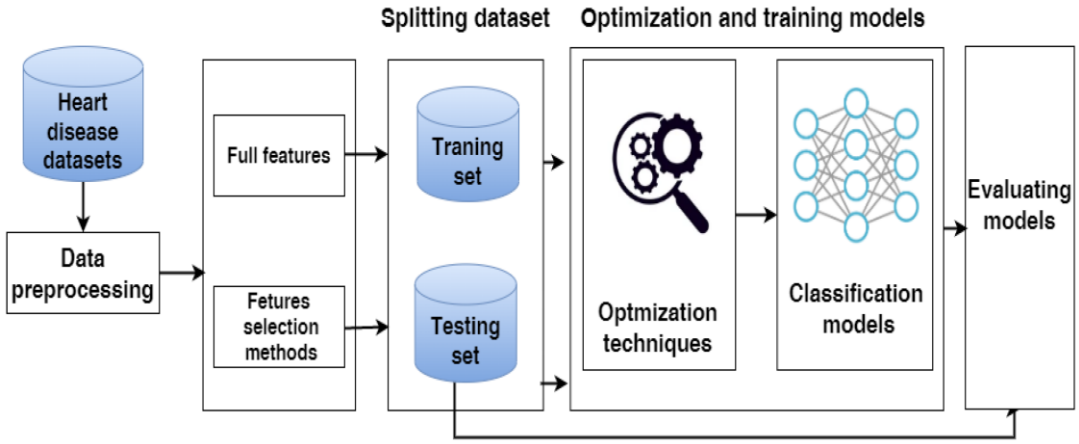

Рисунок 2.1 – Фази прогнозування серцевих захворювань

2.1 Набори даних про серцеві захворювання

У нашій роботі ми використовували два набори даних про серцеві захворювання.

2.1.1 Набір даних 1

Ми використовували великий набір даних про захворювання серця (Heart Disease) [31]. Ці дані включають 18 незалежних ознак та одну залежну змінну як мітку класу для прогнозування захворювань серця. Мітка класу

містить два значення: 0 представляє мітку класу здоров'я, а 1 - мітку класу захворювання серця. У таблиці 2.1 представлено кількість медичних записів для кожного класу в навчальному та тестовому наборах.

Таблиця 2.1 – Кількість медичних записів для кожного класу в наборах даних про серцеві захворювання.

Набір даних	Заняття	Навчальний набір	Набір для тестування	Всього
Набір даних 1	Хвороба серця	21898	5475	27373
	Здоровий	24 000	6000	30000
	Всього	45898	11475	57373
Набір даних Клівленда	Хвороба серця	421	105	526
	Здоровий	399	100	499
	Всього	820	205	1025

2.1.2 Набір даних Клівленда

Набір даних Клівленда [32] містить 13 незалежних змінних як ознаки та одну залежну змінну як мітку класу, що використовується для діагностики серцевих захворювань. Мітка класу містить два значення: 0 представляє мітку класу здорового стану, а 1 представляє мітку класу серцевих захворювань. У таблиці 2.1 представлено кількість медичних записів для кожного класу в навчальних та тестових наборах набору даних Клівленда про серцеві захворювання.

2.2 Попередня обробка даних

Перший набір даних про захворювання серця містить 14 числових

ознак та чотири категоріальні ознаки. Дані були попередньо оброблені після збору наступним чином: видалення дублікатів записів та кодування категоріальних даних у числові дані, такі як куріння та рак шкіри.

2.3 Розділення даних

Два набори даних розділені на два набори за допомогою методу стратифікованої вибірки: 80% навчальних наборів та 20% тестових наборів. Моделі навчаються та оптимізуються за допомогою навчальних даних. Тестовий набір використовується для оцінки та тестування моделі. Метод стратифікованої вибірки – це один із способів розділення набору даних, який використовується для отримання вибірок, що точно відображають розподіл класів у популяції. Він розділяє набір даних на однорідні підмножини; кожна підмножина містить однаковий відсоток кожного класу [33, 34]. Цей метод використовувався в дослідженнях різних галузей охорони здоров'я [35-37].

2.4 Методи вибору ознак

У нашій роботі ми використовуємо метод відбору ознак рекурсивним виключенням ознак (RFE) для вилучення найбільш інформативних ознак з кожного набору даних. RFE визначає основні ознаки, визначаючи високу кореляцію між ознаками та цільовим об'єктом [38] [38]. Він призначає одне значення як ранжування для ознак, якщо ознаки мають високий рівень співпраці з цільовим об'єктом. Нещодавно була представлена нова стратегія RFE, яка використовує RF та SVM для оцінки ознак, а не ефективності класифікації, та вибирає другорядні значущі ознаки для видалення [39,40].

У нашій роботі було застосовано метод рекурсивного виключення ознак як один із найбільш ефективних підходів для зменшення розмірності вхідних даних і зосередження моделі на найважливіших ознаках. Основна ідея методу полягає у поступовому виключенні найменш значущих ознак,

оцінюючи їхній вплив на якість моделі. На кожній ітерації модель навчається з поточним підмножиною ознак, після чого визначаються ознаки з найменшим вкладом, які потім видаляються. Цей процес повторюється доти, доки не буде досягнуто заданої кількості ознак або поки продуктивність моделі не стабілізується.

RFE ефективно працює в комбінації з різними базовими моделями, які використовуються для оцінювання важливості ознак. У нашій роботі в якості таких моделей було використано Random Forest (RF) та Support Vector Machine (SVM).

Random Forest (RF) – ансамблевий метод на основі дерев рішень, який оцінює важливість ознак за рахунок середнього зменшення імпурності (наприклад, критерію Джині або ентропії) в кожному дереві. Це дозволяє враховувати взаємозв'язки між ознаками та виявляти ті, що найчастіше впливають на кінцевий результат класифікації. Перевагою RF є його стійкість до перенавчання і здатність працювати з великою кількістю ознак, включаючи корельовані.

Support Vector Machine (SVM) – модель, що визначає гіперплощину, яка найкраще розділяє класи. При використанні в RFE, коефіцієнти ваг (у випадку лінійного ядра) відображають ступінь важливості ознак. Ознаки з найменшими вагами виключаються на кожному кроці. Такий підхід дозволяє виявити ознаки, які мають найбільший вплив на розмежування класів, що особливо важливо в задачах з обмеженою кількістю прикладів або у високовимірних просторах.

У нашому дослідженні RFE дозволив відібрати ключові ознаки для кожного з наборів даних, зменшити розмірність простору ознак, покращити швидкість обчислень і, головне, підвищити точність класифікації. Комбінування RFE з RF і SVM забезпечило збалансовану та обґрунтовану оцінку інформативності кожної ознаки, дозволяючи уникнути включення зайвих або слабо значущих параметрів у модель.

2.5 Підхід машинного навчання

2.5.1 Алгоритми машинного навчання

Ми протестували багато класичних моделей машинного навчання з різних сімейств, включаючи SVM [41-44], логістичну регресію (LR) [45,46], баєсівський теорем Нейва (NB) [47], дерево рішень (DT) [48], випадковий ліс (RF) [49, 50] та K-найближчих сусідів (k-NN) [51].

2.5.2 Методи оптимізації для класичних моделей

Пошук за сіткою використовується для точного налаштування гіперпараметрів різних класичних моделей машинного навчання шляхом генерації дискретних сіток у межах області гіперпараметрів та вибору списку параметрів, які забезпечують найкращу продуктивність [52]. Дані розділяються на два сегменти за допомогою методу перехресної перевірки: один використовується для навчання та перевірки моделей (навчальний набір), а інший використовується для тестування моделі (тестовий набір) [19]. Навчальний набір використовувався для перевірки моделей за допомогою методу k-кратної перехресної перевірки.

2.6 Гібридні моделі

2.6.1 Гібридні архітектури моделей

Ми запропонували дві гібридні моделі: CNN-LSTM та CNN-GRU для прогнозування серцевих захворювань. Структури гібридних моделей проілюстровано на рисунку 2.2

Рисунок 2.2 – Архітектура гібридних моделей CNN-LSTM та CNN-GRU, що використовуються для прогнозування серцевих захворювань.

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.

2.6.2 Методи оптимізації для гібридних моделей

Для оптимізації гібридних моделей використовується баєсівський

оптимізатор. Цей метод пошуку швидко генерує простір пошуку та знаходить найкращі значення гіперпараметрів для моделей [53]. Ми використовуємо налаштування параметрів для CNN-LSTM та CNN-GRU, як показано в таблиці 2.2.

Таблиця 2.2 – Встановлення значень параметрів

Параметри	Значення
filters	[16,128]
Kernel_size	[2,3,4,5]
Pool_Size	[2,3,4,5]
Unit_LSTM	від 20 до 500
Unit_GRU	від 20 до 500
Unit_Dense	від 20 до 500

2.7 Запропонована модель ансамблю стекування

У цій роботі наша модель розроблена з використанням двох рівнів: Рівень 1 та Рівень 2, як показано на рисунку 2.3. 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.

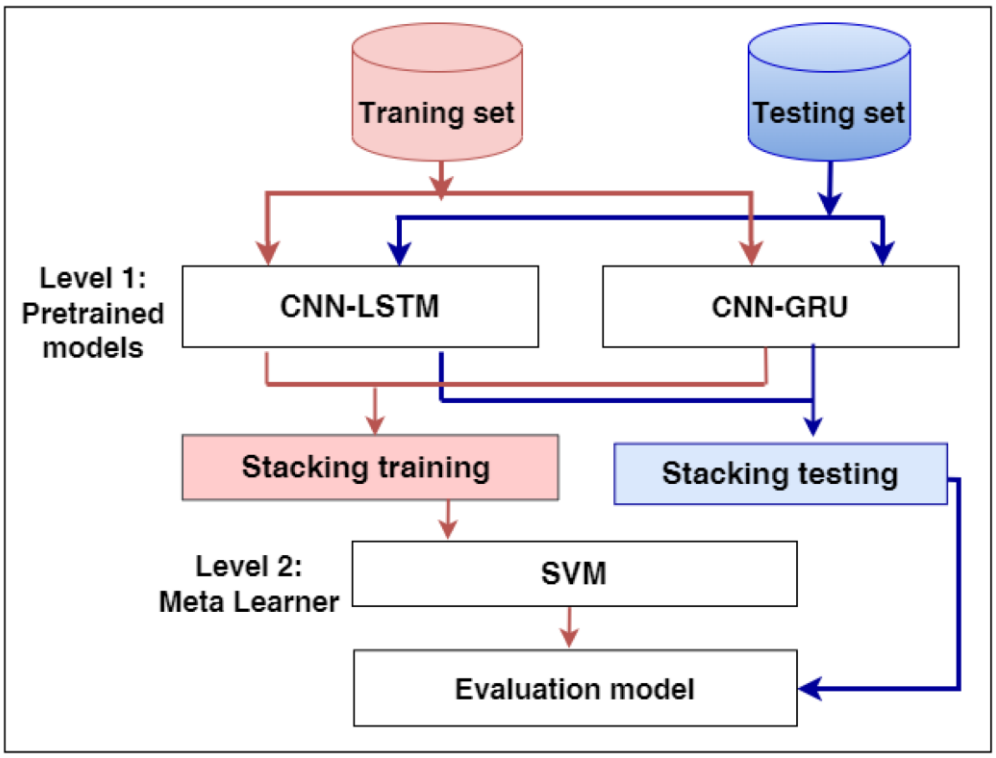

Рисунок 2.1 – Запропонована модель прогнозування серцевих захворювань

2.8 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).

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

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

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

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

3 ТЕСТУВАННЯ ІНФОРМАЦІЙНОЇ СИСТЕМИ

У цьому розділі ми описуємо ранг ознак після застосування RFE до двох наборів даних. Крім того, ми описуємо результати використання моделей машинного навчання (SVM, LR, RF, NB та KNN), гібридних моделей (CNN-LSTM, CNN-GRU) та запропонованої моделі для повних та вибраних ознак.

3.1 Експериментальна установка

Експерименти в цій роботі реалізовано за допомогою Google Colab з бібліотеками Python, такими як Scikit-learn, TensorFlow та іншими. Ми використовували пошук за сіткою та баєсівський оптимізатор для оптимізації моделей машинного навчання та гібридних моделей. Ми використовували метод RFE для визначення найкращих ознак з двох наборів даних. Два набори даних розділено на два набори: 80% навчальний та 20% тестовий набір з використанням стратифікованих методів. Моделі навчаються та тестуються за допомогою навчального та тестового наборів відповідно.

3.2 Результати набору даних 1

3.2.1 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 3.1. 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.

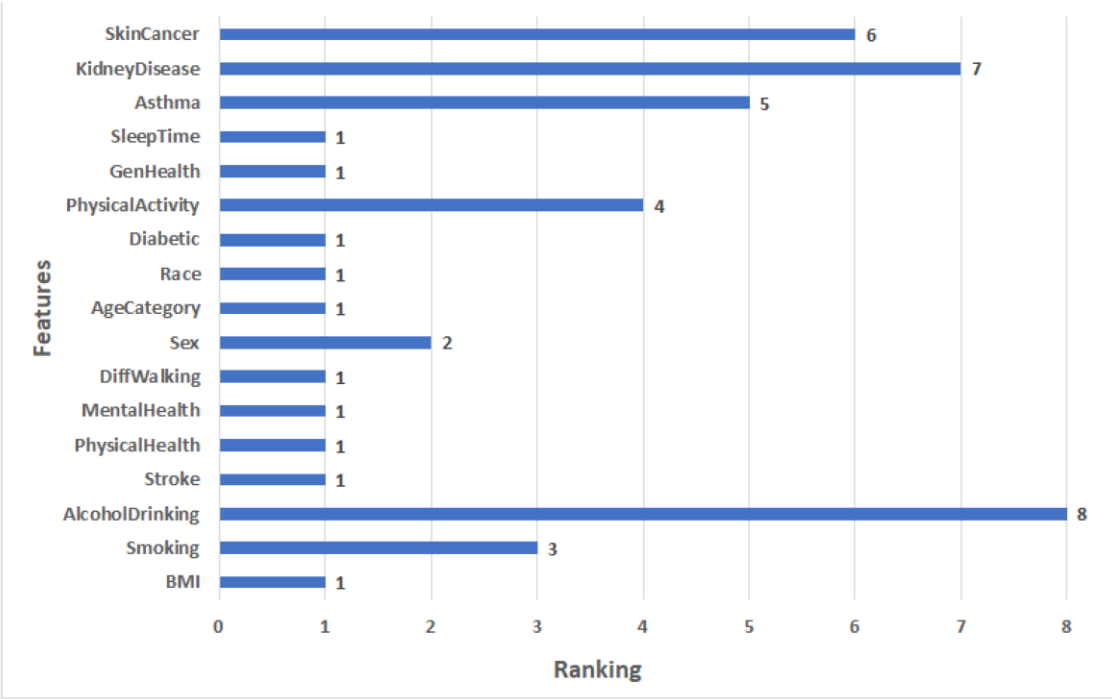

Рисунок 3.1 – Ознаки ранжування для набору даних серцевих захворювань 1

3.2.2 Результати застосування моделей

У цьому розділі 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, наведено в таблиці 3.1

В таблиці 3.2 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.

Таблиця 3.1 – Найкращі значення параметрів для CNN-LSTM та CNN-GRU.

Набір даних	Моделі	Параметри	Повні функції	Вибрані функції
Dataset 1	CNN-LSTM	filters	128	16
		Kernel_size	4	4
		Pool_Size	2	2
		Unit_LSTM	380	40
		Unit_Dense	140	50
	CNN-GRU	filters	128	16
		Kernel_size	4	4
		Pool_Size	2	2
		Unit_GRU	100	320
		Unit_Dense	100	200

Таблиця 3.2 – Результат застосування моделей з повним набором функцій та вибраних функцій для набору даних 1.

Підходи	Моделі	Особливості	Матриця продуктивності			
			ACC	PRE	REC	F1
1	2	3	4	5	6	7
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

Продовження таблиці 3.2

1	2	3	4	5	6	7
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

Результати повного використання функцій.

Для моделей ML, RF та LR реєструють приблизно однакові найвищі бали (75,32% ACC, 75,44% PRE, 75,32% REC, 75,33% F1) та (75,60% ACC, 75,60% PRE, 75,60% REC, 75,60% F1) відповідно. NB фіксує найгірші бали (60,87% ACC, 64,98% PRE, 60,87% REC, 56,69% F1). KNN реєструє другі найвищі бали (73,16% ACC, 73,47% PRE, 73,16% REC, 73,16% F1).

Для гібридних моделей CNN-LSTM має найвищі бали (76,64% ACC, 76,9% PRE, 76,64% REC та 76,65% F1). CNN-GRU фіксує найнижчі бали (75,63% ACC, 75,65% PRE, 75,63% REC, 75,58% F1).

Запропонована модель реєструє найвищі бали (ACC = 78,81%, PRE 78,1%, REC 78,81% та F1 78,81%) порівняно з іншими моделями. Вона покращує ACC на 2,17, PRE на 1,2, REC на 2,17 та F1 на 2,16 порівняно з CNN-LSTM.

Результати вибраних функцій.

Для моделей ML, RF та LR реєструють приблизно однакові найвищі бали (73,02% ACC, 73,06% PRE, 73,02% REC, 73,03% F1) та (73,58% ACC, 73,60% PRE, 73,58% REC, = 73,59% F1) відповідно. NB реєструє найгірші бали (60,84% ACC, 64,97% PRE, 60,84% REC, F1 = 56,63%). KNN реєструє другі найвищі бали (72,59% ACC, 72,92% PRE, 72,59% REC, F1 = 72,59%).

Найвищі бали для гібридних моделей належать CNN-LSTM (75,22% ACC, 75,42% PRE, 75,22% REC та 75,22% F1). Найнижчі бали зафіксовано у

CNN-GRU (74,07% ACC, 74,23% PRE, 74,07% REC та 74,08% F1).

Порівняно з іншими моделями, запропонована модель досягає найвищих балів (77,42% ACC, 77,99% PRE, 77,42% REC та 77,39% F1). Порівняно з CNN-LSTM, вона покращує ACC на 2,2%, PRE на 2,57%, REC на 2,2% та F1 на 2,17%.

3.3 Результати набору даних Клівленда

3.3.1 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 3.2. 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.

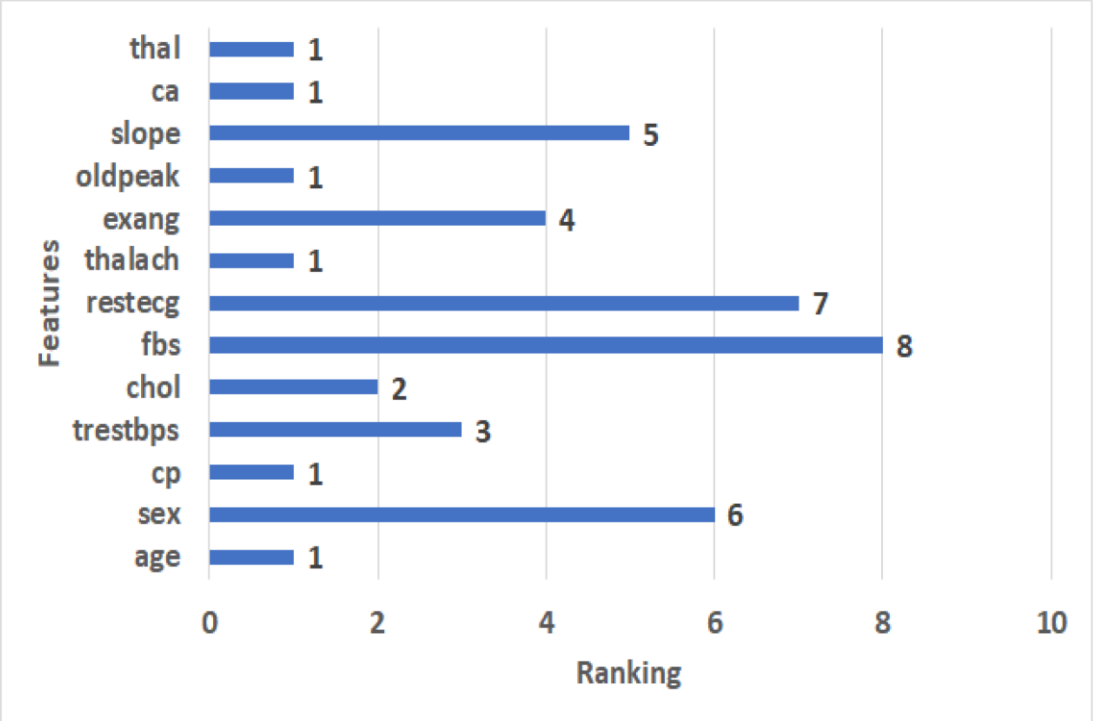

Рисунок 3.2 – Ознаки ранжування для набору даних Клівленда

3.3.2 Результати застосованих моделей

У цьому розділі представлено налаштування значень параметрів для моделей та результати застосованого машинного навчання, гібридних моделей та запропонованої моделі з повними та вибраними ознаками для набору даних Cleveland. Для гібридних моделей CNN-LSTM та CNN-GRU було змінено такі налаштування: розмір пакета = 50, епоха = 50, швидкість навчання = 0,00004, а як оптимізатор використовується Adam. Деякі з найкращих значень гіперпараметрів CNN-LSTM та CNN-GRU, визначені KerasTuner, наведено в таблиці 3.3

Таблиця 3.3 – Найкращі значення параметрів для набору даних Клівленда.

Набори даних	Моделі	Параметри	Повні функції	Вибрані функції
Cleveland dataset	CNN-LSTM	filters	128	16
		Kernel_size	4	5
		Pool_Size	2	2
		Unit_LSTM	360	60
		Dense Unit	160	20
	CNN-GRU	filters	64	16
		Kernel_size	4	5
		Pool_Size	2	2
		Unit_GRU	440	80
		Unit_Dense	160	40

У таблиці 3.4 наведено результати застосування машинного навчання, гібридних моделей та запропонованої моделі з повним набором функцій та вибраними функціями за допомогою RFE до набору даних Клівленда.

Таблиця 3.4 – Результат застосування моделей з повними та вибраними ознаками для набору даних Клівленда.

Підходи	Моделі	Особливості	Матриця 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	67.32	67.43	67.3	67.18
		Вибрані функції	73.17	73.19	73.17	73.14
	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	86.34	86.41	86.34	86.34
	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

Повні функції.

Для моделей ML, RF має найвищі бали (86,34% ACC, 86,34% PRE, 86,34% REC та 86,34% F1). NB фіксує найнижчі бали (60,00% ACC, 60,06% PRE, 60,00% REC, 59,74% F1). DT реєструє другі за величиною бали (82,44% ACC, 82,46% PRE, 82,44% REC, 82,44% F1).

Для гібридних моделей CNN-LSTM має найвищі бали (89,76% від ACC, 89,96% від PRE, REC = 89,76% від REC, F1 = 89,75%). CNN-GRU фіксує найнижчі бали (88,29% від ACC, 89,06% від PRE, REC = 88,29% від REC, 88,26% від F1).

Запропонована модель реєструє найвищі бали (97,17% ACC, 97,41% PRE, 97,17% REC, 97,16% F1) порівняно з іншими моделями. Вона покращує ACC на 7,41, PRE на 7,46, REC на 7,41 та F1 на 7,4 порівняно з CNN-LSTM.

Вибрані функції.

Для моделей ML, RF має найвищі бали (82,93% ACC, 82,99% PRE, 82,93% REC, 82,92% F1). NB фіксує найнижчі бали (64,88% ACC, 64,90% PRE, 64,88% REC, 64,88% F1). DT реєструє другі за величиною бали (81,95% ACC, PRE = 82,01%, 81,95% REC, 81,93% F1).

Для гібридних моделей CNN-LSTM має найвищі бали (86,34% ACC, 86,41% PRE, 86,34% REC та 86,34% F1). CNN-GRU фіксує найнижчі бали (85,85% ACC, 86,92% PRE, 85,85% REC, 85,78% F1).

Запропонована модель реєструє найвищі бали (91,22% ACC, 91,29% PRE, 91,22% REC, 91,23% F1) порівняно з іншими моделями. Вона покращує ACC на 4,88, PRE на 4,88, REC на 4,88 та F1 на 4,88 порівняно з CNN-LSTM.

3.4 Аналіз результатів

Ми використали два набори даних про захворювання серця, завантажені з Kaggle. Ми застосували методи відбору ознак RFE для вибору основних ознак. Запропонована модель у всіх випадках досягла найвищого балу порівняно з іншими моделями.

3.4.1 Набір даних 1

На рисунках 3.3 та 3.4 показано найкращі моделі для застосування моделей з повним набором функцій та вибраними функціями. Ми бачимо, що запропонована модель досягла найвищих балів з повним набором функцій при ACC = 78,81%, PRE = 78,81%, REC = 78,81% та F1 = 78,81% порівняно з іншими моделями з повним набором функцій та вибраними функціями, і вона покращує ACC на 2,17, PRE на 1,2, REC на 2,17 та F1 на 2,16 порівняно з CNN-LSTM. Крім того, вона має найвищі бали з вибраними функціями при (ACC = 77,42%, PRE = 77,99%, REC = 77,42%, F1 = 77,39%), і вона покращує ACC на 2,2%, PRE на 2,57%, REC на 2,2% та F1 на 2,17%. LR має найнижчі бали з повним набором функцій та вибраними функціями.

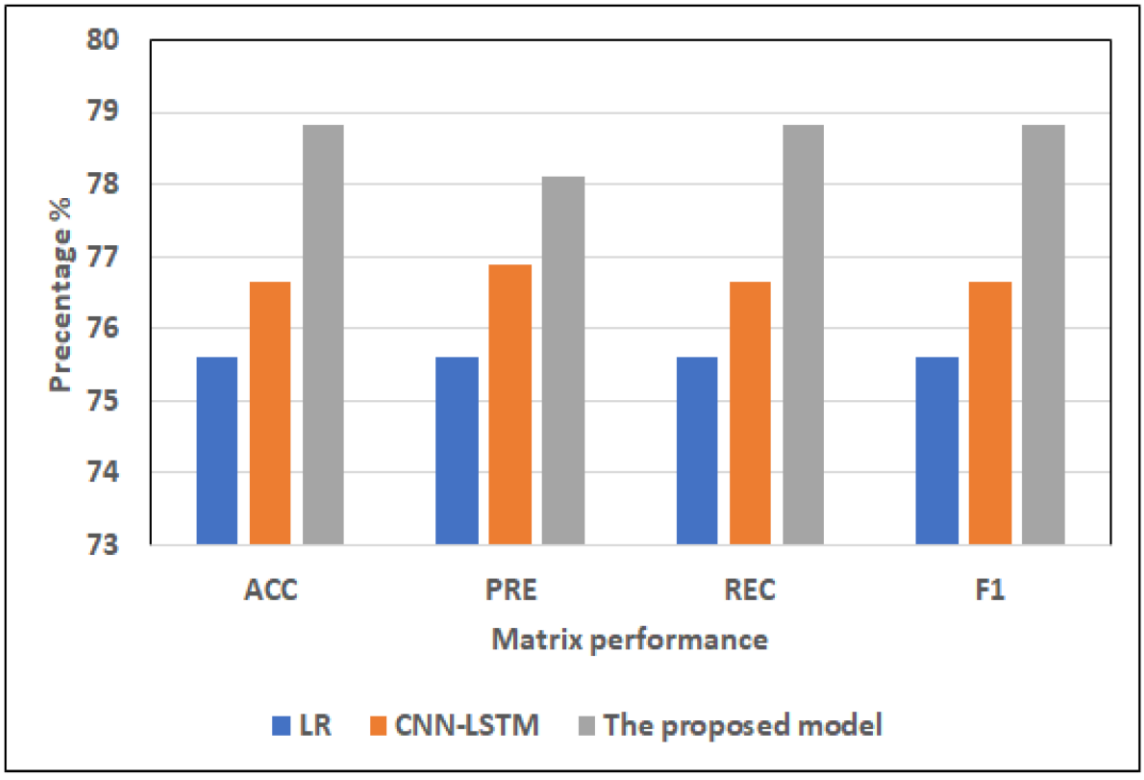

Рисунок 3.3 – Найкращі моделі для застосування моделей з повним набором функцій для набору даних 1.

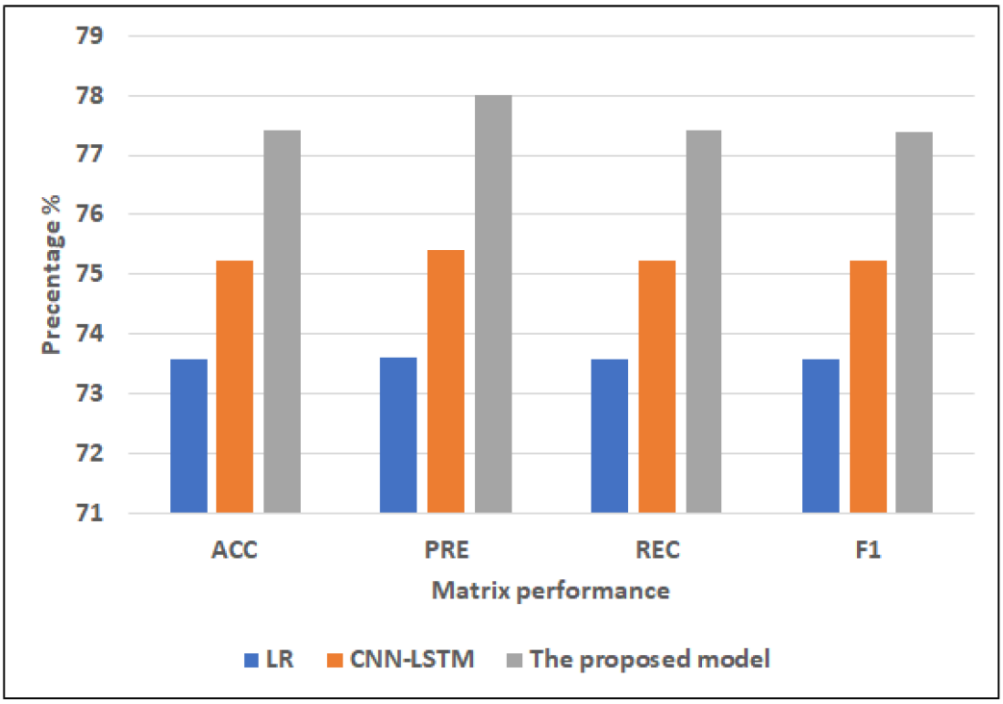

Рисунок 3.4 – Найкращі моделі для застосування моделей з вибраними ознаками для набору даних 1.

3.4.2 Набір даних Клівленда

На рисунках 3.5 та 3.6 показано найкращі моделі для застосування моделей з повним набором функцій та вибраними функціями. Ми бачимо, що запропонована модель досягла найвищих балів з повним набором функцій при ACC = 98,17%, PRE = 98,42%, REC = 98,17% та F1 = 98,15% порівняно з іншими моделями з повним набором функцій та вибраними функціями, і вона покращує ACC на 3,41, PRE на 3,46, REC на 3,41 та F1 на 3,4 порівняно з CNN-LSTM. Крім того, він має найвищі бали за вибраними ознаками при (ACC = 91,22%, PRE = 91,29%, REC = 91,22%, F1 = 91,22%), і покращує ACC на 4,88, PRE на 4,88, REC на 4,88 та F1 на 4,88 порівняно з CNN-LSTM. RF має найнижчі бали за повним набором ознак, а LR має найнижчі бали за вибраними ознаками.

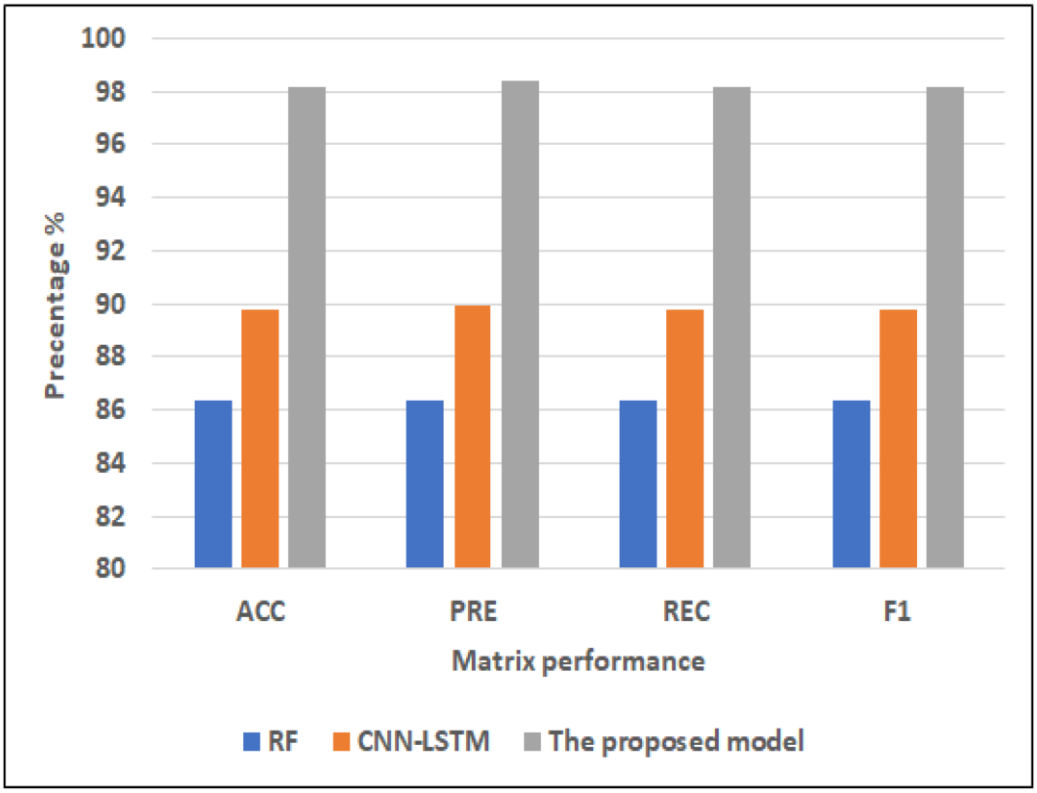

Рисунок 3.5 – Найкращі моделі для застосування моделей з повним набором функцій для набору даних 2.

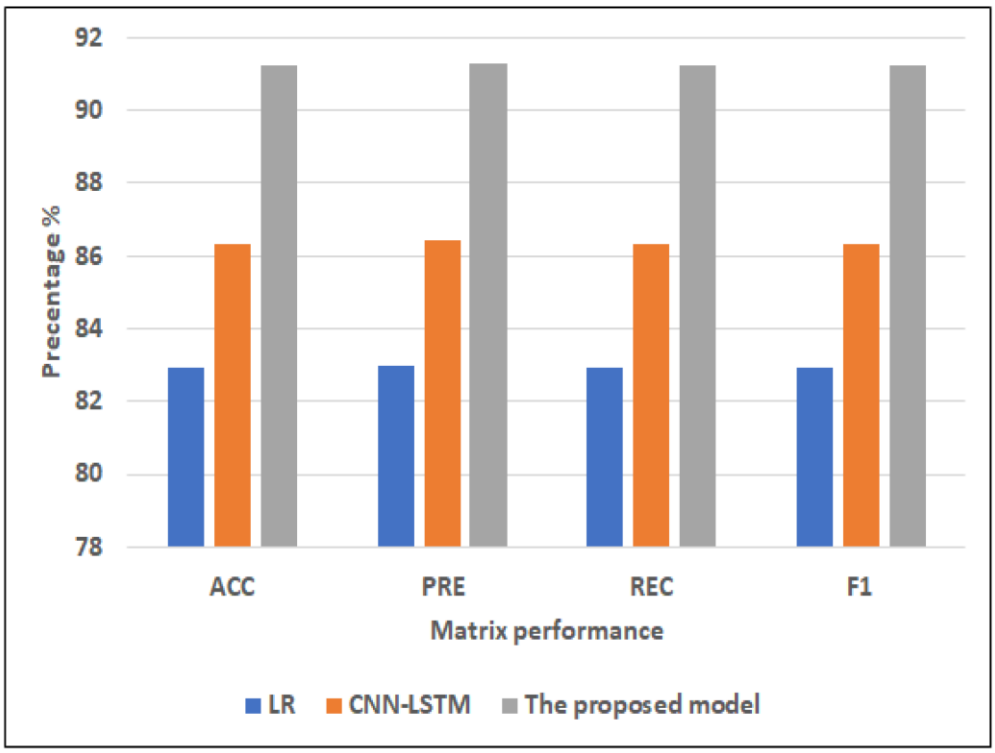

Рисунок 3.6 – Найкращі моделі для застосування моделей з вибраними ознаками для набору даних 2.

ВИСНОВКИ

У кваліфікаційній роботі запропоновано ансамбль глибокого стекінгу інформаційної системи для покращення ефективності прогнозування серцевих захворювань. Запропонована модель базувалася на інтеграції двох попередньо навчених та оптимізованих глибоких гібридних моделей: CNN-LSTM та CNN-GRU. Класифікатор SVM використовувався як модель мета-навчання. Першою гібридною моделлю була модель CNN-LSTM, яка поєднувала шари CNN та LSTM. Другою гібридною моделлю була модель CNN-GRU, яка поєднувала моделі CNN з GRU. RFE використовувався для вибору найважливіших ознак з двох наборів даних про серцеві захворювання. Запропоновані моделі порівнювали з п'ятьма класичними моделями машинного навчання, включаючи LR, RF, K-NN, DT, NB та гібридні моделі (тобто CNN-LSTM та CNN-GRU). Результати збирали з повним набором ознак та вибраним набором ознак. Порівняно з іншими моделями, результат, отриманий за допомогою запропонованої моделі, мав оптимальну продуктивність з усіма ознаками. Для першого набору даних запропонована модель мала найвищий показник ACC 78,81%, PRE 78,1%, REC 78,81% та F1 78,81. Для набору даних Клівленда запропонована модель мала найвищий ACC 97,17%, PRE 97,42%, REC 97,17% та F1 97,15%. Крім того, запропонована модель отримала кращі результати, ніж літературні. Як результат, запропонована модель може покращити прогнозування захворювань та покращити якість життя пацієнтів із серцевими захворюваннями.

Це робить модель корисною для застосування в реальних клінічних умовах, де точність і надійність мають критичне значення. Використання методів глибокого навчання в поєднанні з оптимізацією ознак дозволяє ефективніше виявляти приховані патерни в медичних даних. Модель також може бути адаптована до інших типів захворювань з подібною структурою

даних. Подальші дослідження можуть бути зосереджені на її впровадженні у вигляді медичного програмного забезпечення або мобільного застосунку для підтримки діагностики. Таким чином, робота має як наукову, так і практичну цінність.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Cardiovascular Diseases (CVDs). Available online: http://www.who.int/cardiovascular_diseases/en/ .

2. Hall, J.E.; Hall, M.E. Guyton and Hall Textbook of Medical Physiology e-Book; Elsevier Health Sciences: Amsterdam, The Netherlands, 2020.

3. Bhowmick, A.; Mahato, K.D.; Azad, C.; Kumar, U. Heart Disease Prediction Using Different Machine Learning Algorithms. In Proceedings of the 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), Sonbhadra, India, 17–19 June 2022; pp. 60–65.

4. Saleh, H.; Alyami, H.; Alosaimi, W. Predicting Breast Cancer Based on Optimized Deep Learning Approach. Comput. Intell. Neurosci. 2022, 2022, 1820777.

5. Cardoso, M.R.; Santos, J.C.; Ribeiro, M.L.; Talarico, M.C.R.; Viana, L.R.; Derchain, S.F.M. A metabolomic approach to predict breast cancer behavior and chemotherapy response. Int. J. Mol. Sci. 2018, 19, 617.

6. Spagnuolo, G.; De Vito, D.; Rengo, S.; Tatullo, M. COVID-19 outbreak: An overview on dentistry. Int. J. Environ. Res. Public Health 2020, 17, 2094.

7. Alouffi, B.; Alharbi, A.; Sahal, R.; Saleh, H. An Optimized Hybrid Deep Learning Model to Detect COVID-19 Misleading Information. Comput. Intell. Neurosci. 2021, 2021, 9615034.

8. Mitchell, T.; Buchanan, B.; DeJong, G.; Dietterich, T.; Rosenbloom, P.; Waibel, A. Machine learning. Annu. Rev. Comput. Sci. 1990, 4, 417–433.

9. Chan, S.R.; Torous, J.; Hinton, L.; Yellowlees, P. Mobile tele-mental health: Increasing applications and a move to hybrid models of care. Healthcare 2014, 2, 220–233.

10. Sharma, S.; Parmar, M. Heart diseases prediction using deep learning neural network model. Int. J. Innov. Technol. Explor. Eng. (IJITEE) 2020, 9, 124–137.

11. Weissler, E.H.; Naumann, T.; Andersson, T.; Ranganath, R.; Elemento, O.; Luo, Y.; Freitag, D.F.; Benoit, J.; Hughes, M.C.; Khan, F.; et al. The role of machine learning in clinical research: Transforming the future of evidence generation. Trials 2021, 22, 1–15.

12. Melin, P.; Monica, J.C.; Sanchez, D.; Castillo, O. Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: The case of Mexico. Healthcare 2020, 8, 181.

13. Sagi, O.; Rokach, L. Ensemble learning: A survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2018, 8, e1249.

14. Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. ICML Citeseer 1996, 6, 148–156.

15. Rajagopal, S.; Kundapur, P.P.; Hareesha, K.S. A stacking ensemble for network intrusion detection using heterogeneous datasets. Secur. Commun. Netw. 2020, 2020, 4586875.

16. Bühlmann, P. Bagging, boosting and ensemble methods. In Handbook of Computational Statistics; Springer: Berlin/Heidelberg, Germany, 2012; pp. 985–1022.

17. Kavitha, M.; Gnaneswar, G.; Dinesh, R.; Sai, Y.R.; Suraj, R.S. Heart disease prediction using hybrid machine learning model. In Proceedings of the 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 20–22 January 2021; pp. 1329–1333.

18. Ishaq, A.; Sadiq, S.; Umer, M.; Ullah, S.; Mirjalili, S.; Rupapara, V.; Nappi, M. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE Access 2021, 9, 39707–39716.

19. Ansarullah, S.I.; Saif, S.M.; Kumar, P.; Kirmani, M.M. Significance of visible non-invasive risk attributes for the initial prediction of heart disease using different machine learning techniques. Comput. Intell. Neurosci. 2022, 2022, 9580896.

20. Spencer, R.; Thabtah, F.; Abdelhamid, N.; Thompson, M. Exploring

feature selection and classification methods for predicting heart disease. Digit. Health 2020, 6, 2055207620914777.

21. Bharti, R.; Khamparia, A.; Shabaz, M.; Dhiman, G.; Pande, S.; Singh, P. Prediction of heart disease using a combination of machine learning and deep learning. Comput. Intell. Neurosci. 2021, 2021, 8387680.

22. Gokulnath, C.B.; Shantharajah, S. An optimized feature selection based on genetic approach and support vector machine for heart disease. Clust. Comput. 2019, 22, 14777–14787.

23. Amin, M.S.; Chiam, Y.K.; Varathan, K.D. Identification of significant features and data mining techniques in predicting heart disease. Telemat. Informatics 2019, 36, 82–93.

24. Bashir, S.; Khan, Z.S.; Khan, F.H.; Anjum, A.; Bashir, K. Improving heart disease prediction using feature selection approaches. In Proceedings of the 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 8–12 January 2019; pp. 619–623.

25. Javid, I.; Ghazali, R.; Zulqarnain, M.; Husaini, N.A. Deep Learning GRU Model and Random Forest for Screening Out Key Attributes of Cardiovascular Disease. In International Conference on Soft Computing and Data Mining; Springer: Berlin/Heidelberg, Germany, 2022; pp. 160–170.

26. Chae, M.; Gil, H.W.; Cho, N.J.; Lee, H. Machine Learning-Based Cardiac Arrest Prediction for Early Warning System. Mathematics 2022, 10, 2049.

27. Narmadha, S.; Gokulan, S.; Pavithra, M.; Rajmohan, R.; Ananthkumar, T. Determination of various deep learning parameters to predict heart disease for diabetes patients. In Proceedings of the 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 3–4 July 2020; pp. 1–6.

28. Adhikari, B.; Shakya, S. Heart Disease Prediction Using Ensemble Model. In Proceedings of Second International Conference on Sustainable Expert Systems; Springer: Berlin/Heidelberg, Germany, 2022; pp. 857–868.

29. Javid, I.; Alsaedi, A.K.Z.; Ghazali, R. Enhanced accuracy of heart

disease prediction using machine learning and recurrent neural networks ensemble majority voting method. Int. J. Adv. Comput. Sci. Appl. 2020, 11.

30. Ghosh, P.; Azam, S.; Jonkman, M.; Karim, A.; Shamrat, F.J.M.; Ignatious, E.; Shultana, S.; Beeravolu, A.R.; De Boer, F. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access 2021, 9, 19304–19326.

31. Heart Disease Prediction. Available online: https://www.kaggle.com/code/andls555/heart-disease-prediction/data/.

32. Heart Disease Dataset. Available online: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset .

33. Sechidis, K.; Tsoumakas, G.; Vlahavas, I. On the stratification of multi-label data. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases; Springer: Berlin/Heidelberg, Germany, 2011; pp. 145–158.

34. Liberty, E.; Lang, K.; Shmakov, K. Stratified sampling meets machine learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 2320–2329.

35. O. Barkovska, I. Velykodnyi, O. Liashenko, Y. Davydov and I. Ivanisenko, "Study of the Architectural Features of the ResNet Neural Network Model for Solving the Task of Speaker Recognition," *2024 IEEE 5th KhPI Week on Advanced Technology (KhPIWeek)*, Kharkiv, Ukraine, 2024, pp. 1-5, doi: 10.1109/KhPIWeek61434.2024.10877951.

36. Prusty, S.; Patnaik, S.; Dash, S.K. SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. Front. Nanotechnol. 2022, 4, 972421.

37. Fonarow, G.C.; Adams, K.F.; Abraham, W.T.; Yancy, C.W.; Boscardin, W.J.; Committee, A.S.A. Risk stratification for in-hospital mortality in acutely decompensated heart failure: Classification and regression tree analysis. JAMA 2005, 293, 572–580.

38. Srinivasan, B.; Pavya, K. Feature selection techniques in data mining: A

study. Int. J. Sci. Dev. Res. (IJSDR) 2017, 2, 594–598.

39. Saeys, Y.; Inza, I.; Larranaga, P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007, 23, 2507–2517.

40. Liu, H.; Yu, L. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng. 2005, 17, 491–502.

41. Noble, W.S. What is a support vector machine? Nat. Biotechnol. 2006, 24, 1565–1567. [Google Scholar] [CrossRef]

42. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. IEEE Intell. Syst. Their Appl. 1998, 13, 18–28.

43. Pisner, D.A.; Schnyer, D.M. Support vector machine. In Machine Learning; Elsevier: Amsterdam, The Netherlands, 2020; pp. 101–121.