Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет інформаційно-аналітичних технологій та менеджменту
(повна назва)

Кафедра прикладної математики
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Класифікація та виявлення аномалій часових рядів
за допомогою візуалізації рекурентних діаграм
(тема)

Виконав:
студент _2_ курсу, групи _САУм-20-1
Степаненко Ю.С.
(прізвище, ініціали)
Спеціальність
124 Системний аналіз
(код і повна назва спеціальності)
Тип програми _освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма
Системний аналіз і управління
(повна назва освітньої програми)
Керівник ____ проф. Кіріченко Л.О.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ПМ _____ Тевяшев А.Д.
(підпис) (прізвище, ініціали)

2021 р.

Харківський національний університет радіоелектроніки

Факультет _інформаційно-аналітичних технологій та менеджменту
Кафедра _прикладної математики
Рівень вищої освіти _другий (магістерський)
Спеціальність _124 Системний аналіз
(код і повна назва)
Тип програми _освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма _Системний аналіз і управління
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри ПМ _____
(підпис)
"____" _____ 2021 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові _Степаненко Юлії Сергіївні
(прізвище, ім'я, по батькові)

1. Тема роботи _Класифікація та виявлення аномалій часових рядів за допомогою візуалізації рекурентних діаграм

затверджена наказом по університету від 05 листопада 2021 р. № 1642 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 10 грудня 2021 р.

3. Вихідні дані до роботи _датасет, що складається з рядів ЕКГ

4. Перелік питань, що потрібно опрацювати в роботі
1. Системний аналіз проблеми порівняльного аналізу методів класифікації часових рядів на прикладі електрокардіограми
2. Вибір і обґрунтування методу розв'язання
3. Програмна реалізація
4. Результати обчислювального експерименту
5. Аналіз можливих застосувань

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

1. Актуальність теми роботи
2. Постановка задачі
3. Системний аналіз предметної області
4. Метод чисельного аналізу
5. Результати обчислювального експерименту

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Підбір та вивчення технічної літератури за темою роботи	8 – 14 листопада 2021 р.	виконано
2	Вибір та обґрунтування методу	15 – 21 листопада 2021 р.	виконано
3	Розробка алгоритму і програми	22 – 28 листопада 2021 р.	виконано
4	Проведення аналітичних досліджень та розрахунків	29 листопада – 5 грудня 2021 р.	виконано
5	Робота над текстом пояснювальної записки	6 – 9 грудня 2021 р.	виконано
6	Представлення роботи на рецензію в ЕК	10 грудня 2021 р.	виконано

Дата видачі завдання _8_ листопада 2021 р.

Студент __
(підпис)

Керівник роботи _____ проф. Кіріченко Л.О.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 67 с., 7 табл., 22 рис., 1 дод., 14 джерела.

АНОМАЛІЇ, ЕЛЕКТРОКАРДІОГРАМА (ЕКГ), КЛАСИФІКАЦІЯ, МАШИННЕ НАВЧАННЯ, НЕЙРОННІ МЕРЕЖІ, РЕКУРЕНТНІ ДІАГРАМИ, ЧАСОВІ РЯДИ.

Об’єкт дослідження – часові ряди, отримані під час проведення ЕКГ.

Мета роботи – провести класифікацію часових рядів та виявлення аномалій за допомогою візуалізації рекурентних діаграм.

Методи дослідження – метод класифікації часових рядів та пошуку аномалій, заснований на побудові рекурентних діаграм за допомогою методів машинного навчання.

У кваліфікаційній роботі аналізується метод виявлення аномалій та класифікації часових рядів, заснований на візуалізації рекурентних діаграм. Часовий ряд ЕКГ перетворюється в рекурентну діаграму, яка є зображенням, після чого для класифікації та пошуку аномалій використовується нейронна мережа. Була побудована згорткова нейронна мережа. Метод був досліджений на медичних часових рядах, що були отримані під час проведення електрокардіографії. Метод, що використовує згорткові нейронні мережі, раніше не застосовувався у цій галузі. Результати показали, що даний метод має досить високу точність класифікації.

ABSTRACT

Introductory note: 67 pages, 7 tables, 22 figures, 1 appendixes, 14 sources.

ANOMALY, CLASSIFICATION, ELECTROCARDIOGRAM (ECG), MACHINE LEARNING, NEURAL NETWORKS, RECCURENCE PLOT, TIME SERIES.

Object of research – time series obtained during electrocardiography.

Purpose of work – to classify time series and detect anomalies by visualizing recurrent diagrams.

Methods of research – a method of classification of time series and finding anomalies based on the construction of recurrent diagrams using machine learning methods.

The certification's work analyzes the method of detecting anomalies and classification of time series, based on the visualization of recurrent diagrams. As a result of recurrence analysis the time series is converted into a recurrent diagram, which is an image, after which a neural network is used to classify and search for anomalies. A convolutional neural network was constructed. The method was investigated on medical time series obtained during electrocardiography. The method with convolutional neural networks has not been used before. The results showed that this method has a fairly high classification accuracy.

ЗМІСТ

ВСТУП

Актуальність теми. Задача класифікації аномалій у часових рядах є однією з найскладніших задач аналізу даних. В останні рокі почали з'являтися дослідження, в яких для класифікації аномалій часових рядів використовується метод побудови та візуалізації рекурентних діаграм.

Аномалії – це зразки даних, які не відповідають чітко визначеному поняттю нормальної поведінки. Важливість виявлення аномалій обумовлена тим фактом, що аномалії даних приводять до значної і дієвої інформації в медичній залузі. Аномалії на рядах ЕКГ можуть вказувати на наявність злоякісних хвороб серця.

Часовий ряд – це зібраний у різні моменти часу статистичний матеріал про значення будь-яких параметрів досліджуваного процесу. Інакше кажучи, впорядкована у часі послідовність значень будь-якого датчика.

Мета і завдання кваліфікаційної роботи. Метою і завданням кваліфікаційної роботи є пошук аномалій та проведення класифікації часових рядів заснованих на ЕКГ, на основі побудови рекурентних діаграм. Для досягнення поставленої мети необхідно виконати наступні завдання:

– розглянути методи визначення аномалій часових рядів;

– розглянути методи класифікації часових рядів;

– ознайомитися з багатошаровою згортковою нейронною мережою для розпізнавання зображень;

– побудувати рукурентні діаграми на основі часових рядів отриманих після проведення ЕКГ;

– побудувати нейронну мережу для розпізнання зображень;

– розробити програмну реалізацію для класифікації аномалій рядів ЕКГ;

– провести аналіз класифікації аномалій часових рядів отриманих під час проведення ЕКГ;

– на основі отриманих даних зробити висновок про проведену роботу.

Об’єктом дослідження є виявлення аномалій серцевих ритмів.

Предметом дослідження є класифікатор на основі багатошарової згорткової нейронної мережі.

Методи дослідження. У кваліфікаційній роботі використовується метод классифікації для визначення аномаій. Цей метод заснований на тому, що нормальна поведінка системи може визначатися одним чи кількома класами. Тоді екземпляр, який не належить жодному з класів, є аномальним. Цей метод використовує підхід "частково з учителем". У кваліфікаційній роботі використовується новий підхід до застосування рекурентних методів для класифікації – розпізнавання аномалій з зображень рекурентних діаграм.

Публікації. Результати, отримані у кваліфікаційній роботі, було представлено на 25-му Міжнародному молодіжному форумі «Радіоелектроніка та молодь у XXI столітті» (м. Харків, 20–22 квітня 2021 р.) [1]. Також результати роботи було опубліковано в журналі «Системні технології» (2021, № 136, С.81–86) [2] та на конференції «5th International Conference on Computational Linguistics and Intelligent System» (2021) [3].

1 СИСТЕМНИЙ АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧ ДОСЛІДЖЕННЯ

1.1 Системний аналіз проблеми порівняльного аналізу методів класифікації часових рядів на прикладі електрокардіограми

1.1.1 Вербальна модель системи

Об'єктом дослідження є проблема виявлення та класифікація аномалій медичних часових рядів, тобто аномалій серцевих ритмів за допомогою електрокардіограм. Метою проведення дослідження є аналіз методу виявлення та класифікації аномалій, за такими критеріями як: точність проведенної класифікації, складність розробки методу, особливості програмної реалізації, можливість використовувати даний метод класифікації на досить великих наборах даних.

Аномалія визначається як елемент, який явно виділяється з набору даних, до якого він належить, та суттєво відрізняється з інших елементів вибірки.

Виявлення аномалій – це знаходження, під час аналізу, рідкісних даних, подій або спостережень, що викликають підозри через істотну відмінність від більшої частини даних. Зазвичай аномальні дані характеризують певний вид проблеми, наприклад шахрайство у банку, медичні проблеми чи помилки у тексті. Аномалії також згадуються як викиди, незвичайності, шуми, відхилення чи винятки.

Задача класифікації аномалій часових рядів є однією з найскладніших завдань інтелектуального аналізу даних. Існує кілька підходів до класифікації часових рядів, більшість з яких засновані на розрахунку різних метрик між часовими рядами [4].

Пошук аномалій у часових рядях, що заснований на класифікації, з точки зору системного аналізу, – це складна система, що має декілька входів та виходів. Окрім потоків інформації, які протікають всередині системи, існують

потоки інформації між системою та зовнішнім середовищем.

Завдання роботи полягає в аналізі методу виявлення та класифікації аномалій, за такими критеріями як точність проведенної класифікації та оцінка продуктивності реалізованої програми.

1.1.2 Морфологічний опис системи

Зовнішнє середовище – це ті умови, що існують у навколишньому середовищі, незалежно від діяльності підприємства. Але вони можуть відчувати на собі вплив організації та впливати на функціонування організації.

Організація отримує із зовнішнього середовища всі основні ресурси, що є необхідними для функціонування, в наслідок постійного обміну з зовнішнім середовищем.

Можливість якісно та швидко виявити у людини відхилення від норми під час обстеження серцевих ритмів – є призначенням моделі.

Далі почнемо розглядати морфологічний опис системи з опису зовнішнього середовища, який наведено на рис. 1.1.

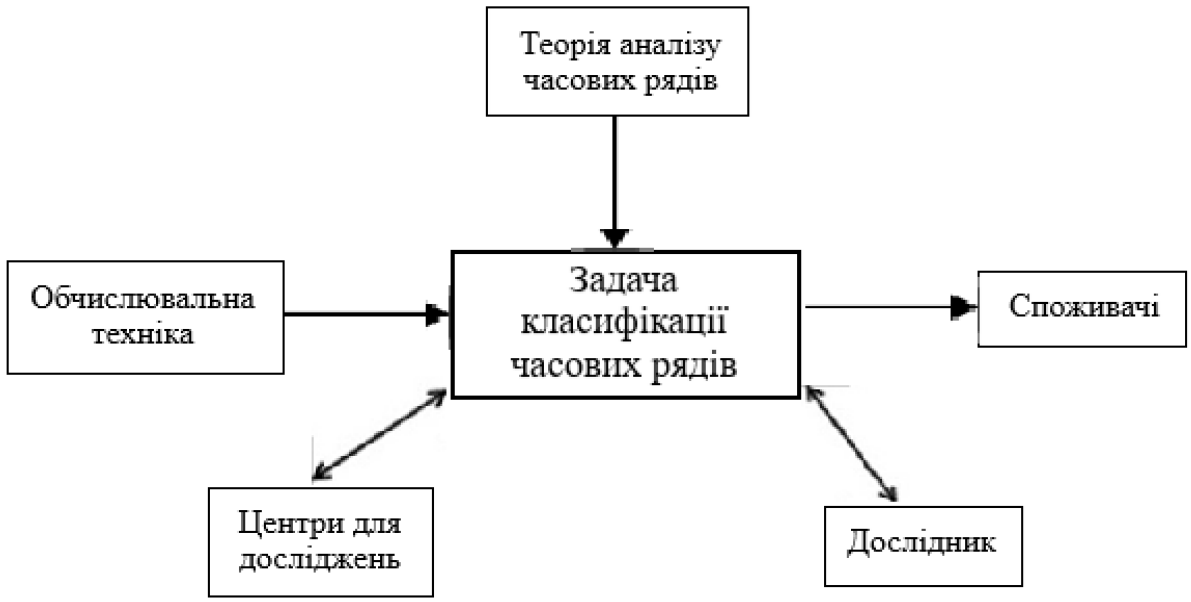


Рисунок 1.1 – Модель зовнішнього середовища системи

Об'єкти зовнішнього середовища:

– медична техніка є засобом для отримання результатів серцевих ритмів;

– обчислювальна техніка для виявлення аномальних імпульсів серцевого ритму з використанням ЕКГ є засобом для отримання практичних результатів ;

– медичні центри для досліджень класифікації аномалій серцевихї ритмів забезпечують умови для проведення обробки цих даних;

– дослідник аналізує данні для класифікації, проводить експерименти, обирає методи виявлення аномалій та способи класифікації часових рядів;

– в ролі споживачів виступають лікарні та інші медичні заклади, підприємства, що займаються дослідженням імпульсів серцевого ритму з використанням ЕКГ;

– на основі теорії методів аналізу часових рядів складаються способи класифікації.

Для опису функціонування моделі із середовищем на рис. 1.2 представлена модель «чорний ящик».

Чорний ящик – це система яку можна розглядати з точки зору її входів та виходів, без інформації про її внутрішню роботу. Ця модель є вихідною при побудові моделі складної системи. Така модель акцентує увагу дослідника на взаємодії системи з зовнішнім середовищем. Система чорний ящик має «вхід», у який поступає інформація, тобто вплив середовища на систему, і «вихід» – це результат роботи, тобто цільові продукти для показу результатів. Стан виходів зазвичай функціонально залежить від стану входів. Іншими словами, коли використовують термін «чорний ящик», мають на увазі систему, яка виконує перетворення інформації, але при цьому користувача не цікавить, як вона це робить.

У даній системі «входом» є медичні часові ряди на основі імпульсів серцевого ритму, які в подальшому будуть аналізовані та класифіковані за допомогою згорткових нейронних мереж. На «вихід» отримуємо виявлення аномалії у часових рядах, та на цих результатах діагностування хвороби.

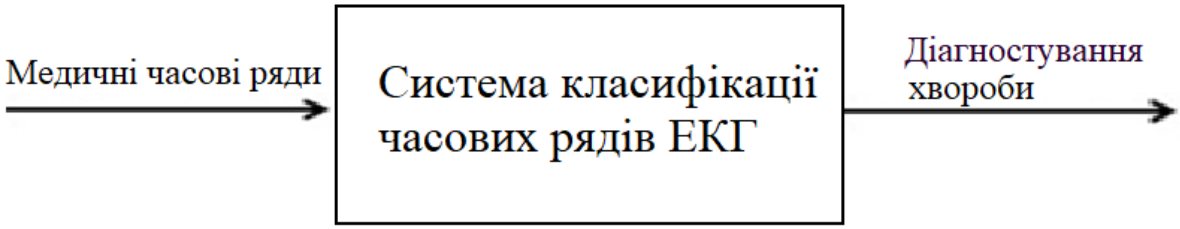


Рисунок 1.2 – Модель «чорний ящик»

1.1.3 Функціональна модель системи

IDEF0 – методологія функціонального моделювання та графічного опису процесів, яка призначена для опису процесів. Для нових систем застосування IDEF0 має за ціль означення вимог та функцій для подальшої розробки системи. Вони відповідають вимогам та реалізують виділені функції. Особливістю IDEF0 є її акцент на ієрархічне представлення об'єктів, що значно полегшує розуміння предметної області. В IDEF0 розглядаються логічні зв'язки між роботами, а не послідовність їх виконання в час.

Перша діаграма в ієрархії діаграм IDEF0 є загальним описом системи та її взаємодії із зовнішнім середовищем. Такі діаграми називають контекстними. Контекстна діаграма зображує функціонування системи в цілому. Після опису системи проводиться її декомпозиція на великі фрагменти. Після декомпозиції контекстної діаграми проводиться декомпозиція кожного великого фрагмента системи на більш дрібні і так далі, до досягнення потрібного рівня деталізації опису. Діаграма IDEF0 представлена на рис. 1.3.

Проводимо декомпозицію системи, для того щоб детально розглянути функціональну частину. Декомпозиція роботи «Класифікація часових рядів» зображена на рис. 1.4. Даний процес розділено на три задачі:

– попередня обробка даних;

– виділення ознак часових рядів;

– проведення аналізу класифікації.

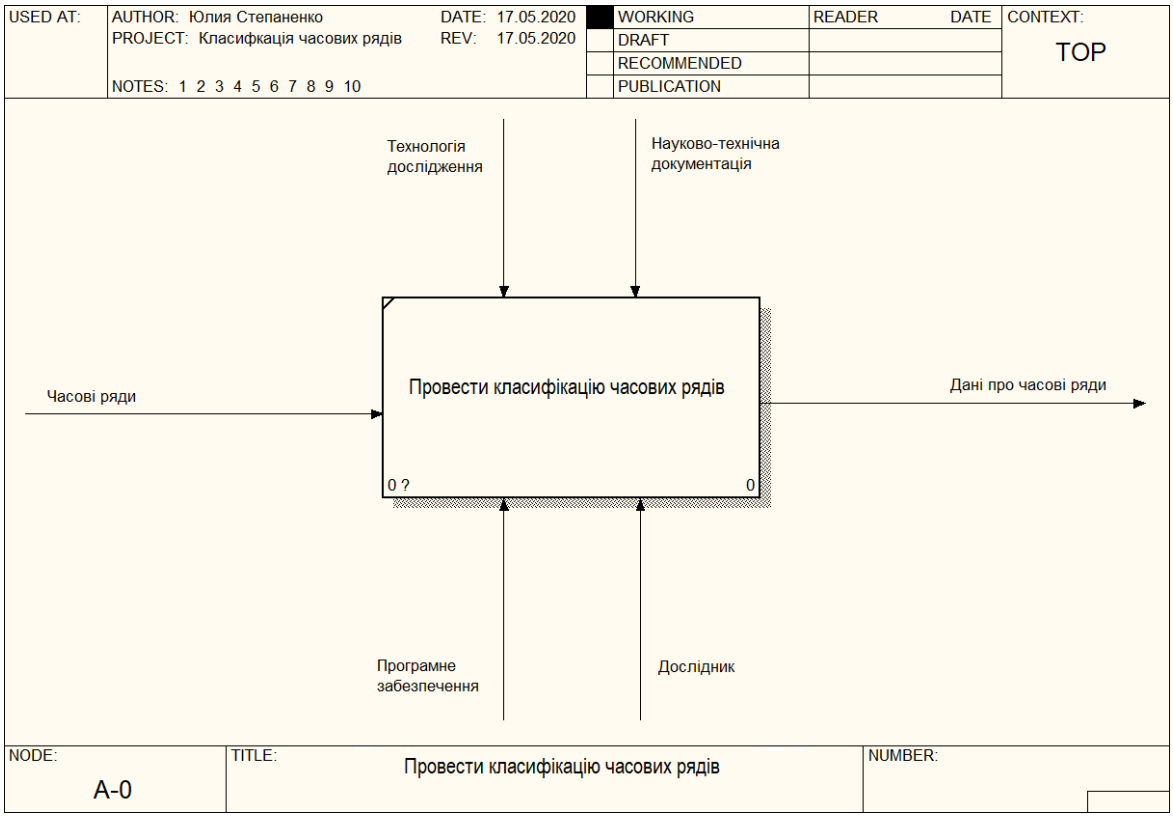


Рисунок 1.3 – Контекстна діаграма IDEF0

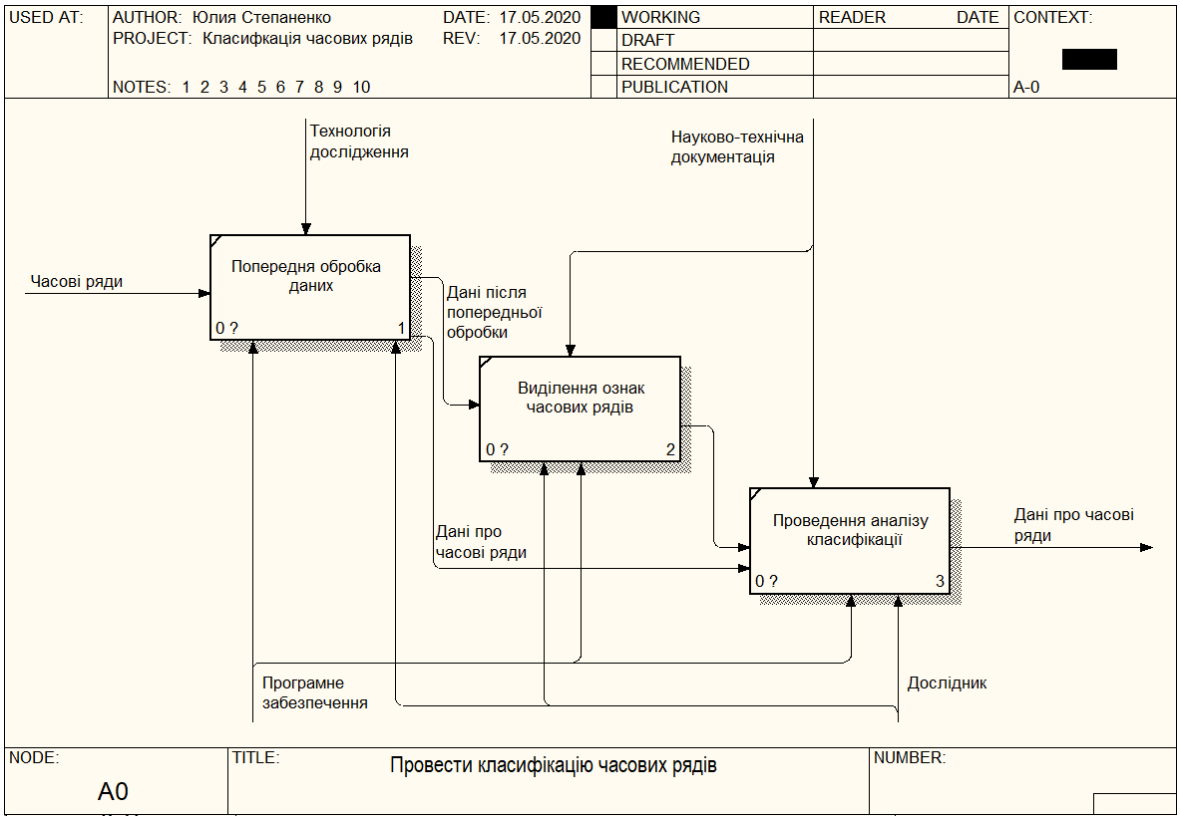


Рисунок 1.4 – Декомпозиція роботи «Провести класифікацію часових рядів»

1.1.4 Інформаційна модель

Інформаційна модель системи відображає зв’язки між елементами системи у вигляді структур даних. Також вона акцентує увагу дослідника на склад та взаємозв’язках потоків даних. Діаграми потоків даних (Data Flow Diagramming, DFD) – це нотація, призначена для моделювання інформаційних систем з точки зору зберігання, обробки та передачі даних. DFD використовуються для обробки інформації.

Подібно до IDEF0, DFD є модельною системою, та вони зображують досліджувану систему у вигляді мережі пов’язаних між собою робіт. Також діаграму потоків даних можна використовувати як додаток до моделі IDEF0 для більш наочного відображення поточних операцій [5]. Діаграма DFD представлена на рис. 1.5.

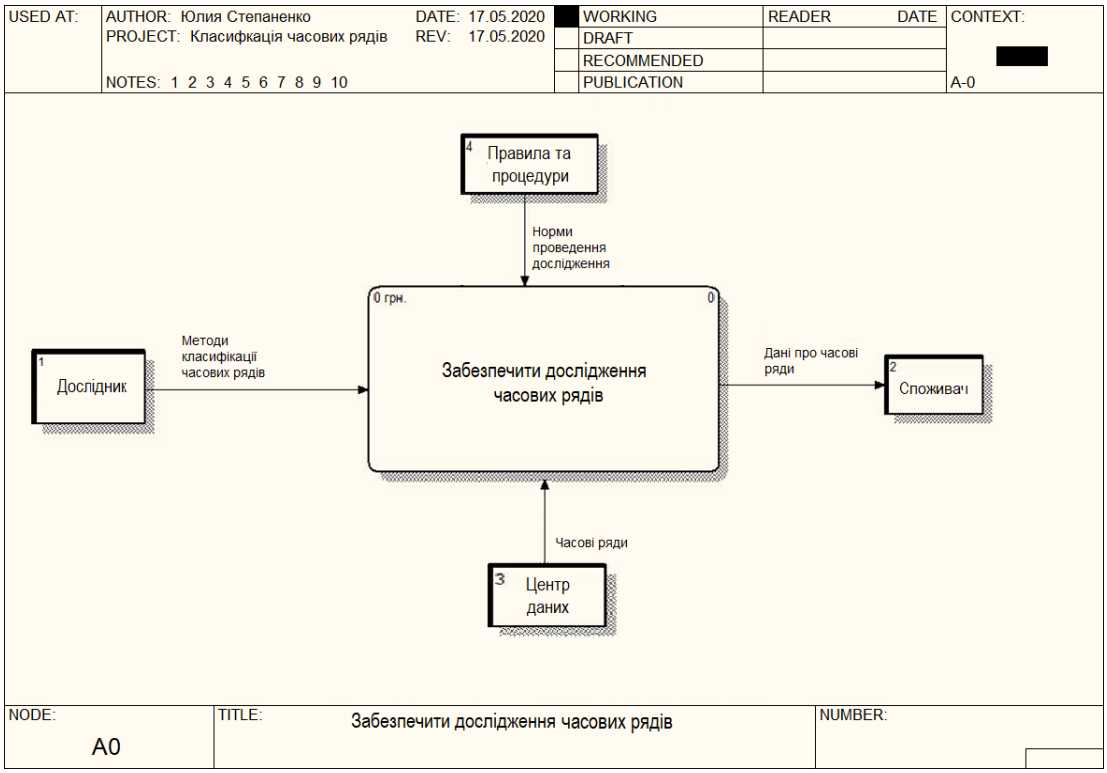


Рисунок 1.5 – DFD-діаграма

1.2 Аналіз сценаріїв вирішення проблеми класифікації медичних часових рядів

1.2.1 Модель аналізу проблеми

Об'єктом дослідження є модель порівняльного аналізу методів виявлення аномалій часових рядів та класифікації на прикладі імпульсів серцевого ритму.

Метою дослідження є знаходження найбільш оптимального та точного методу для класифікації часових рядів задля виявлення аномалій медичного ряду, тобто знаходження хвороби серця у пацієнта.

Математична модель буде розроблена при застосуванні декількох алгоритмів. Алгоритми будуть порівнюватися з точки зору точності висновків та особливостей програмної реалізації. В результаті буде визначено найбільш оптимальний алгоритм, що може бути застосований для вирішення задачі виявлення та класифікації аномалій [6].

Далі наведено критерії, які здійснюють найбільший вплив на очікуваний результат вибору алгоритма:

– надійність (К1);

– універсальність (К2);

– швидкодія (К3);

– ресурсоємкість (К4).

Надалі будемо робити аналіз критеріїв деатльніше та будемо проводити конкретизацію запропонованих альтернативах. Альтернативами будуть виступати методи, які дуже часто використовувались у дослідах для знаходження аномалій у медичних часових рядах.

Порівнюючи альтернативи на швидкодію, перевага віддаватиметься алгоритму, який буде мати найменьшу кількість ітерацій.

Порівнюючи на універсальність, нас буде цікавити алгоритм, що потребує меншу кулькість інформації на вхід.

При порівнянні альтернативи на надійність алгоритмів, перевага

віддаватиметься алгоритму у якому будемо отримувати однозначний результат від роботи класифікатора медичних часових рядів.

Ресурсоємкість альтернатив будуть необхідні для успішного закінчення роботи алгоритмів без помилок. Тобто ми будемо враховувати загальний час виконання розрахунків.

Розглядаємо такі альтернативи:

– метод опорних векторів (А1);

– метод рекурентних діаграм (А2);

– наївний байесовский класифікатор (А3).

На рис. 1.6 зображена ієрархічна модель процесу аналізу незадоволеностей.

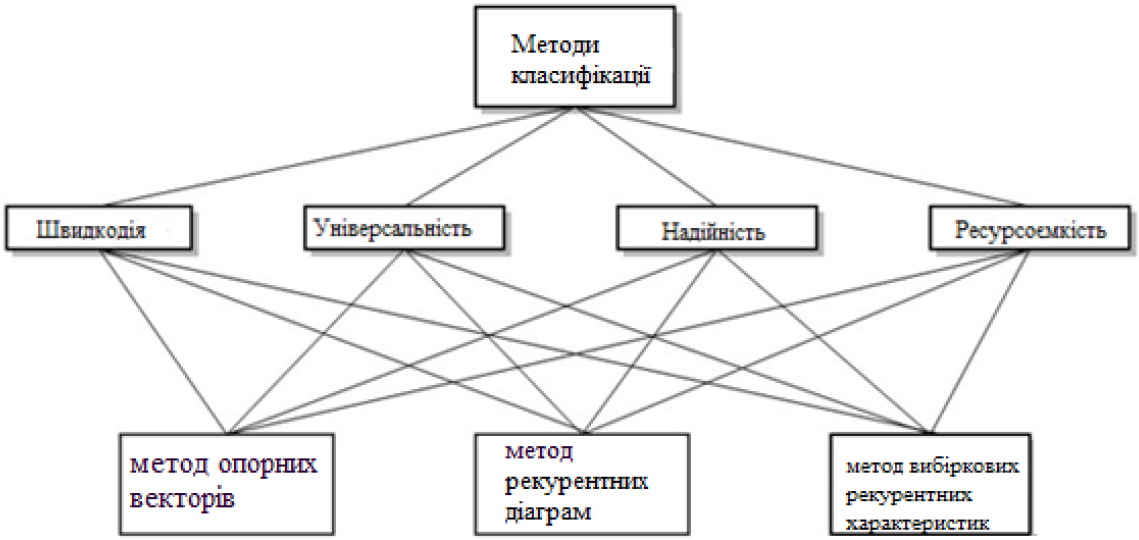


Рисунок 1.6 – Ієрархічна модель процесу аналізу незадоволеностей

1.2.2 Оцінювання вектора пріоритетів незадоволеностей методом аналізу ієрархій

Метод аналізу ієрархій – це структурований метод організації та аналізу складних рішень, заснований на математиці та психології. Він представляє

точний підхід для кількісної оцінки ваги критеріїв прийняття рішень. Для оцінки відносної величини факторів за допомогою парних порівнянь використовується досвід окремих експертів. Кожен з респондентів повинен порівняти відносну важливість між двома пунктами відповідно до спеціально розробленої анкети.

Метод аналізу ієрархій використовується у всьому світі в широкому спектрі ситуацій прийняття рішень у багатьох галузях. Метод аналізу ієрархій допомагає знайти рішення, яке найкраще відповідає цілі та розумінню проблеми [7].

Робимо аналіз невдоволеностей шляхом побудови ієрархічної моделі:

а) нульовий рівень – компоненти проблеми;

б) перший рівень – класифікація незадоволеностей;

в) другий рівень – характеристики компонентів, що впливають на розв’язок поставленої задачі.

Почнемо оцінювати ступінь впливу кожної з груп незадоволеностей на розв’язок проблеми в цілому за допомогою методу парних порівнянь.

На першому рівні аналізу проблеми побудуємо матрицю попарних порівнянь критеріїв. Метою побудови матриці попарних порівнянь є оцінка впливу кожної незадоволеності на поставлену проблему. Результати наведені в табл. 1.1.

Таблиця 1.1 – Матриця попарних порівнянь критеріїв

	К1	К2	К3	К4	Вектор пріоритетів
К1	1,0	7,0	1,0	5,0	0,509
К2	0,143	1,0	3,0	0,2	0,113
К3	1,0	0,333	1,0	5,0	0,238
К4	0,2	5,0	0,2	1,0	0,140

Знаходимо суми елементів матриці за стовбцями для знаходження

індексу узгодженості:

$$y_1 = 1{,}0 + 0{,}143 + 1{,}0 + 0{,}2 = 2{,}343,$$
$$y_2 = 7{,}0 + 1{,}0 + 0{,}333 + 5{,}0 = 13{,}333,$$
$$y_3 = 1{,}0 + 3{,}0 + 1{,}0 + 0{,}2 = 5{,}2,$$
$$y_4 = 5{,}0 + 0{,}2 + 5{,}0 + 1{,}0 = 11{,}2.$$

Тоді

$$\lambda_{\max} \approx 2{,}343 \cdot 0{,}509 + 13{,}333 \cdot 0{,}113 + 5{,}2 \cdot 0{,}238 + 11{,}2 \cdot 0{,}140 = 5{,}505$$

та індекс узгодженості:

$$CI^k = \frac{5{,}505 - 4}{4 - 1} = 0{,}68.$$

Оскільки матриця попарних порівнянь критеріїв є матрицею четвертого порядку, відношення узгодженості буде:

$$CR^k = \frac{CI^k}{0{,}9} = 0{,}187.$$

Матриця попарних порівнянь критеріїв побудована правильно, бо відношення узгодженості є близьким до 0,1.

Далі сформуємо матриці попарних альтернатив за кожним критерієм. Метою є порівняння методів між собою за кожним критерієм окремо.

У табл. 1.2 – 1.5 наведено матриці порівнянь за кожним ж критерієв, які здійснюють найбільший вплив на очікуваний результат вибору алгоритма.

Таблиця 1.2 – Матриця порівнянь за критерієм К1

К1	А1	А2	А3	Вектор пріоритетів
А1	1,0	0,111	0,143	0,051
А2	9,0	1,0	5,0	0,722
А3	7,0	0,2	1,0	0,227

Для знаходження індексу узгодженості знаходимо суми елементів матриці за стовбцями:

$$y_1 = 1{,}0 + 9{,}0 + 7{,}0 = 18{,}0,$$
$$y_2 = 0{,}111 + 1{,}0 + 0{,}2 = 1{,}311,$$
$$y_3 = 0{,}143 + 5{,}0 + 1{,}0 = 6{,}243.$$

Тоді

$$\lambda_{\max} \approx 0{,}051 \cdot 18{,}0 + 10{,}722 \cdot 1{,}311 + 6{,}243 \cdot 0{,}227 = 3{,}282$$

та індекс узгодженості:

$$CI_{K1}^{A} = \frac{3{,}282 - 3}{3 - 1} = 0{,}141.$$

Оскільки матриця попарних порівнянь альтернатив є матрицею третього порядку, відношення узгодженості:

$$CR_{K1}^{A} = \frac{CI^{k}}{0{,}58} = 0{,}241.$$

Для знаходження індексу узгодженості потрібно знайти суми елементів

матриці за стовбцями.

Таблиця 1.3 – Матриця порівнянь за критерієм К2

К2	А1	А2	А3	Вектор пріоритетів
А1	1,0	7,0	7,0	0,773
А2	0,143	1,0	2,0	0,139
А3	0,143	0,5	1,0	0,088

Знаходимо суми елементів матриці за стовбцями:

$$y_1 = 1{,}0 + 0{,}143 + 0{,}143 = 1{,}286,$$
$$y_2 = 7{,}0 + 1{,}0 + 0{,}5 = 8{,}5,$$
$$y_3 = 7{,}0 + 2{,}0 + 1{,}0 = 10{,}0.$$

Тоді

$$\lambda_{\max} \approx 0{,}773 \cdot 1{,}286 + 0{,}139 \cdot 8{,}5 + 0{,}088 \cdot 10{,}0 = 3{,}056$$

та індекс узгодженості:

$$CI_{K2}^{A} = \frac{3{,}056 - 3}{3 - 1} = 0{,}028.$$

Відношення узгодженості:

$$CR_{K2}^{A} = \frac{CI^{k}}{0{,}58} = 0{,}048.$$

Для знаходження індексу узгодженості знаходимо суми елементів.

Таблиця 1.4 – Матриця порівнянь за критерієм К3

К3	А1	А2	А3	Вектор пріоритетів
А1	1,0	0,111	0,2	0,063
А2	9,0	1,0	3,0	0,672
А3	0,333	0,333	1,0	0,265

Знаходимо суми елементів матриці за стовбцями:

$$y_1 = 1{,}0 + 9{,}0 + 0{,}333 = 10{,}333,$$
$$y_2 = 0{,}111 + 1{,}0 + 0{,}333 = 1{,}444,$$
$$y_3 = 0{,}2 + 3{,}0 + 1{,}0 = 4{,}2.$$

Тоді

$$\lambda_{max} \approx 0{,}063 \cdot 10{,}333 + 0{,}672 \cdot 1{,}444 + 0{,}265 \cdot 4{,}2 = 3{,}028$$

та індекс узгодженості:

$$CI_{K3}^{A} = \frac{3{,}028 - 3}{3 - 1} = 0{,}014.$$

Отже, відношення узгодженості дорівнює:

$$CR_{K3}^{A} = \frac{CI^{k}}{0,58} = 0,024.$$

Таблиця 1.5 – Матриця порівнянь за критерієм К4

К4	А1	А2	А3	Вектор пріоритетів
А1	1,0	0,2	0,333	0,105
А2	5,0	1,0	3,0	0,637
А3	3,0	0,333	1,0	0,258

Для знаходження індексу узгодженості знаходимо суми елементів матриці за стовбцями:

$$y_1 = 1{,}0 + 5{,}0 + 3{,}0 = 9{,}0,$$
$$y_2 = 0{,}2 + 1{,}0 + 0{,}333 = 1{,}533,$$
$$y_3 = 0{,}333 + 3{,}0 + 1{,}0 = 4{,}333.$$

Тоді

$$\lambda_{\max} \approx 0{,}105 \cdot 9{,}0 + 0{,}637 \cdot 1{,}533 + 0{,}258 \cdot 4{,}333 = 3{,}039$$

та індекс узгодженості:

$$CI_{K4}^{A} = \frac{3{,}028 - 3}{3 - 1} = 0{,}020.$$

Відношення узгодженості:

$$CR_{K4}^{A} = \frac{CI^{k}}{0{,}58} = 0{,}034.$$

Для розрахунку вектору глобальних пріоритетів знаходимо добуток:

$$\vec{p} = \begin{bmatrix} 0,051 & 0,773 & 0,063 & 0,105 \\ 0,722 & 0,139 & 0,672 & 0,637 \\ 0,227 & 0,088 & 0,256 & 0,58 \end{bmatrix} \cdot \begin{bmatrix} 0,509 \\ 0,113 \\ 0,238 \\ 0,140 \end{bmatrix} = \begin{bmatrix} 0,143 \\ 0,632 \\ 0,225 \end{bmatrix}.$$

Знайдемо індекс узгодженості, відношення узгодженості для всієї ієрархії:

$$CI = 0,168 + 0,509 \cdot 0,141 + 0,113 \cdot 0,048 + 0,238 \cdot 0,024 + 0,140 \cdot 0,020 = 0,254,$$

$$RI = 0,90 + 0,58 = 1,48,$$

$$CR = \frac{CI}{RI} = \frac{0,254}{1,48} = 0,172,$$

що теж можна вважати доброю узгодженістю.

Отже, найбільша компонента вектора локальних пріоритетів критеріїв відповідає першому критерію. Тоді наступні пріоритети за критеріями порівняння: швидкодія, надійність, ресурсоємність, універсальність. Найбільша компонента вектору глобальних пріоритетів відповідає другій альтернативі – методу рекурентних діаграм. Надалі будемо використовувати метод рекурентних діаграм під час експерименту для виявлення аномалій та класифікації часових рядів.

1.3 Змістовна та формальна постановка задачі

1.3.1 Змістовна постановка задачі

Інфаркти та інсульти зазвичай є гострими захворюваннями та відбуваються, головним чином, внаслідок серцево-судинних захворювань,

таких як аритмія, інфаркт міокарда, стенокардія, тахікардія, ішемія. Такі захворювання досить добре піддаються лікуванню, якщо їх виявити на ранній стадії.

Активність клітин серцевого м'яза фіксується за допомогою електрокардіографії. Подібний аналіз оцінює функціональний стан серця та дозволяє визначити наявні в ньому патології. Електрокардіографія дає можливість з великою точністю говорити про локалізацію вогнищевих змін міокарда, їх розповсюдженість, глибину і час появи.

Електрокардіографія (скорочено ЕКГ) – метод графічної реєстрації електричних явищ, що виникають у серцевому м'язі під час його діяльності, з поверхні тіла. Криву, яка відображає електричну активність серця, називають електрокардіограмою (ЕКГ). Таким чином, ЕКГ – це запис коливань різниці потенціалів, які виникають у серці під час його збудження.

ЕКГ дозволяє виявити аномальні процеси в міокарді, що виникають під впливом різних токсичних речовин. Поєднання електрокардіографічного дослідження з функціональними пробами допомагає виявити приховану коронарну недостатність, перехідні порушення ритму, проводити диференційний діагноз між функціональними та органічними порушеннями роботи серця.

Важливим кроком у визначенні хвороби серця є виявлення аномалій коливать серцевого ритму та класифікація серцевих скорочень. Ритм сигналу ЕКГ можна визначити, якщо знати якою є класифікація серцевих скорочень в сигналах.

Класифікація скорочень може бути дуже трудомістким процесом. Отже, будь-яка автоматизована обробка ЕКГ, що допомагає цьому процесу, є у центрі уваги дослідження.

ЕКГ реєструється на міліметровій сітці, що дозволяє виконати вимірювання частоти серцевих скорочень, тривалості та амплітуди окремих морфологічних елементів запису. Це і є данні на вхід класифікатора. При стандартній швидкості руху електрографічної стрічки 25 мм/с, проміжок часу

між тонкими вертикальними лініями сітки становить 0,04 секунди, а між товстими лініями становить 0,2 секунди.

Стандартна ЕКГ – це запис електричних потенціалів у 12 відведеннях. ЕКГ складається з біполярних відводів від кінцівок I, II і III, однополярних відведеннях від кінцівок aVR, aVL і aVF, та однополярних грудних відводів V1, V2, V3, V4, V5 і V6, та називається 12-свинцева ЕКГ. Кожен свинець – це вид електричної активності серця з певного кута по всьому тілу. Свинець II – найбільш часто використовуваний для смуги ритму. Він дає огляд найбільш важливих хвиль: P, Q, R, S і T.

На рис. 1.7 можемо бачити, що серцевий цикл розділений на зубці та інтервали. Кожен удар серця містить серію відхилень від базової лінії на ЕКГ, хвилі, що відображають еволюцію часу електричної активності в серці. Відхилення вгору чи вниз від базової лінії – хвилі P, Q, R, S, T, U. Зубці Q, R, S мають назву комплекс QRS (без R = комплекс QS). Горизонтальна лінія між зубцями U та P або між зубцями T та P, якщо зубці U не виявляються – це ізоелектрична лінія (ізолінія). Фрагменти лінії між зубцем P та комплексом QRS, а також між комплексом QRS та зубцем T — це сегменти PQ та ST. Фрагменти кривої, що складаються з сегмента та сусіднього зубця, називаються інтервалами PQ та QT.

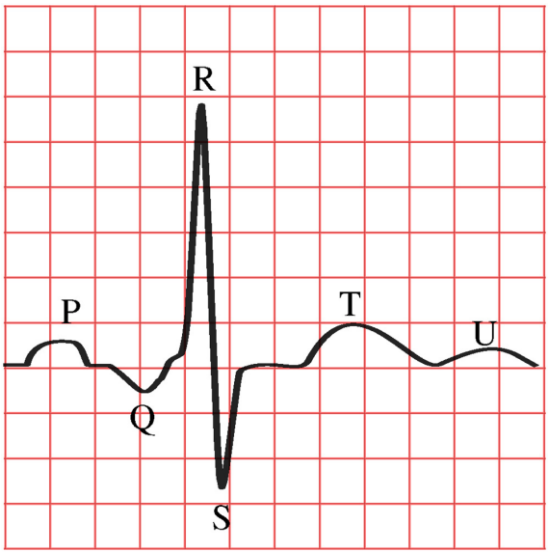


Рисунок 1.7 – Хвилі свинцевої ЕКГ

Пошук аномалій коливать у рядах ЕКГ дозволяє виявити взаємозв'язок між характеристиками і видом захворювання. Класифікація дозволяє запропонувати найефективніший спосіб діагностування хвороб серця. Тому, виконання даної роботи є актуальним для медичних закладів.

У цій роботі ми будемо класифікувати медичні часові ряди, а саме ряди ЕКГ, за допомогою методів машинного навчання. Докладніше розгорнуті питання у пункті 2.2.

1.3.2 Формальна постановка задачі класифікації

Якщо є часткова інформація про аномалії, тобто коли відома нормальна поведінка системи, то можна вирішити її як задачу класифікації з «нормальною» поведінкою та «аномальною».

Задача класифікації – це математична задача, що має безліч об'єктів, деяким чином розподілених на класи.

Класифікація використовується для навчання моделі даних, віднесених до різних класів та віднесення екземплярів даних до одного з наявних класів з використанням отриманої моделі. Методи виявлення аномалій, засновані на класифікації, припускають, що якщо класифікатор, може бути навчений у наявному просторі ознак, то він зможе розділити нормальні та аномальні об'єкти.

Метою задачі є побудова алгоритму, що класифікує певний довільний об'єкт із заданої множини. Тобто надати об'єкту свій номер чи ім’я класу, до якого він належить.

Для задачі виявлення аномалій зазвичай є нормальний опис роботи системи. При навчанні таких даних потрібно побудувати модель нормальної роботи системи, яка у подальшому могла б пророкувати, чи є поточна ситуація на об'єкті «нормальною» або «аномальною».

У роботі розглядається задача виявлення аномалій під час часткового навчання з учителем [8]. Передбачається, що у навчальній вибірці є лише приклади «нормальних» об'єктів. Відповідні методи будують модель, що описує нормальну поведінку системи, і використовують її виявлення аномалій в тестових даних.

В формальному виді постановка задачі класифікації має такий вигляд. Нехай $x_i \in X$, $i = \overline{1,n}$, – множина об’єктів ознак, входів моделі, $y_i \in Y$, $i = \overline{1,n}$, – множина об’єктів відповідей, виходів моделі. Пара $(x_i, y_i) \in X \times Y$ називається розмічений об’єкт, або прецедент.

Кінцева множина $\{x_i\}$, $i = \overline{1,n}$, представляє собою матрицю $\{x_{i,j}\}$, $i = \overline{1,n}$ $j = \overline{1,m}$, розміром $n \times m$, де рядок матриці – це масив ознак одного об’єкта, $\{y_i\}$, $i = \overline{1,n}$, – вектор відповідей, елемент якого є значення номеру класу.

Комбінація $\{x_i\}$, $i = \overline{1,n}$, та $\{y_i\}$, $i = \overline{1,n}$, називається навчальною вибіркою. Задача класифікації полягає у визначенні функції залежності $f : X \to Y$, яка пророкує по $x \in X$ відповіді $y \in Y$.

1.4 Постановка задач дослідження

Метою дослідження є виявлення аномалій медичних часових рядів та проведення аналізу класифікації часових рядів, отриманих за допомогою електрокардіограми, на основі рекурентних діаграм. Сформулюємо задачі для дослідження в рамках даної кваліфікаційної роботи:

– розглянути методи визначення аномалій часових рядів;

– розглянути методи класифікації часових рядів;

– ознайомитися з багатошаровою згортковою нейронною мережою для розпізнавання зображень;

– побудувати рекурентні діаграми на основі часових рядів отриманих після проведення ЕКГ;

– побудувати нейронну мережу для розпізнання зображень;

– розробити програмну реалізацію для класифікації аномалій рядів ЕКГ;

– провести аналіз класифікації аномалій часових рядів отриманих під час проведення ЕКГ;

– на основі отриманих даних зробити висновок про проведену роботу.

2 ВИБІР ТА ОБҐРУНТУВАННЯ МЕТОДА РОЗВ'ЯЗАННЯ

2.1 Рекурентний аналіз як метод аналізу часових рядів

Рекурентний аналіз є методом нелінійної динаміки, що використовуються для аналізу часових рядів, та для виявлення неочевидних залежностей в динаміці ряду. Рекурентний аналіз часових реалізацій базується на фундаментальних властивостях дисипативних динамічних систем – рекурентності, тобто повторюваність станів.

Цю властивість було відзначено ще в 80-х роках століття французьким математиком Пуанкаре і згодом сформульовані у вигляді «теореми рекурентності». Якщо система зводить свою динаміку в обмежену підмножину фазового простору, то система майже напевно, тобто з імовірністю, майже рівний 1, як завгодно близько повертається до якого-небудь спочатку заданого режиму.

Суть цієї фундаментальної властивості в тому, що не дивлячись на те, що навіть найменше обурення в складній системі може привести систему до експоненціального відхилення від її стану, через деякий час система прагне повернутися в стан, деяким чином близького до попереднього, і проходить при цьому подібні етапи еволюції [9].

Рекурентна діаграма є проекцією m-вимірного псевдофазового простору на площину. Нехай точка $x(i)$ відповідає i-й точці фазової траєкторії, що описує обрану динамічну систему у m-мірному просторі, для $i = 1, ..., N$, тоді рекурентний графік є масивом точок $N \times N$, де ненульовий елемент з координатами (i, j) відповідає випадку, коли $x(j)$ достатньо близько до $x(i)$.

Рекурентність стану в момент i при різних значеннях часу j відтворюється всередині двомірної квадратної матриці з чорними і білими точками, де чорні точки позначають наявність рекурентності, обидві координатні осі є осями часу. Таке представлення було назване рекурентною

діаграмою (recurrence plot, RP), оскільки воно фіксує інформацію про рекурентну поведінку системи.

Математично вищесказане описується як

$$R_{i,j} = \Theta\left(\varepsilon_i - \left\| x_i - x_j \right\|\right),\ x_i \in R^m,\ i, j = 1,...N,$$

де N – кількість даних станів;

x_i, ε_i – розмір околиці точки $\vec{x}$ у момент i;

$\|\cdot\|$ – норма;

$\Theta(\cdot)$ – функція Хевісайда.

Непрактично та, як правило, неможливо знайти повну рекурентність у значенні $x_i \equiv x_j$ (стан динамічної, а особливо – хаотичної системи не повторюється повністю еквівалентно початковому стану, а підходить до нього скільки завгодно близько). Таким чином, рекурентність визначається як достатня близькість стану x_j до стану x_i. Іншими словами, рекурентними є стани x_j, які потрапляють в m-вимірну околицю з радіусом ε_i і центром в x_i. Ці точки x_j називаються рекурентними точками (recurrence points).

Оскільки $R_{i,i} = 1, i = 1,..., N$, за визначенням, то рекурентна діаграма завжди містить чорну діагональну лінію – лінію ідентичності (line of identity, LOI) під кутом $\frac{\pi}{4}$ до осей координат. Довільно узята рекурентна точка не несе якої-небудь корисної інформації про стани в часи i і j. Тільки вся сукупність рекурентних точок дозволяє відновити властивості системи [10].

Зовнішній вигляд рекурентної діаграми дозволяє отримати уявлення про характер процесів, які відбуваються в системі, існування і вплив шуму, станів повторення і стабільності, про здійснення в ході еволюції системи різких змін стану.

2.2 Методи пошуку аномалій засновані на класифікації

Аномалія визначається як елемент, який явно виділяється з набору даних, до якого він належить, та суттєво відрізняється від інших елементів вибірки. Задача виявлення аномалій ставиться як задача пошуку в наборах даних зразків, що не задовольняють деякій передбачуваній типовій поведінці.

Для задачі виявлення аномалій зазвичай є нормальний опис роботи системи – наприклад, набір станів системи, у яких кількість збоїв мінімальна. Опис ситуацій, що відповідають неполадкам на об'єкті, часто відсутній чи є неповним. При навчанні таких даних потрібно побудувати модель нормальної роботи системи, яка у подальшому могла б пророкувати, чи є поточна ситуація на об'єкті «нормальною» або «аномальною», тобто чи присутні в даний момент якісь несправності чи ні.

Залежно від наявності або відсутності міток даних виділяють три категорії методів виявлення аномалій:

– виявлення аномалій «з учителем» (методи керованого виявлення аномалій): для методів, що відносяться до цієї категорії, потрібна наявність у навчальній вибірці об'єктів, що відносяться як до нормальних, так і до аномальних. На підставі таких даних будується модель, яка зможе визначати клас об'єктів, що подаються на вхід. Для успішного функціонування подібних моделей необхідна репрезентативна вибірка. Крім того, даних, що належать до «нормальних», зазвичай набагато більше ніж «аномальних», і за рахунок цього створюється дисбаланс, який може вплинути на здатність моделі точно відносити об'єкти до правильних класів;

– виявлення аномалій «без учителя»: для даної категорії методів передбачається, що дані навчання не потрібні. Але при цьому робиться припущення про те, що «нормальні» об'єкти зустрічаються набагато частіше, ніж «аномальні». Інакше методи страждають від великої кількості хибних спрацьовувань;

– виявлення аномалій при частковому навчанні з «учителем» – це

середнє між першими двома: передбачається, що у навчальній вибірці є лише приклади «нормальних» об'єктів. Відповідні методи будують модель, що описує нормальну поведінку системи, і використовують її виявлення аномалій в тестових даних. Також існує невелика кількість методів, для яких вихідними даними є набір «аномалій». Дані методи не знайшли широкого застосування, тому що дуже важко отримати вичерпний набір даних, що описує аномальну поведінку [11].

Зазвичай, задача пошуку аномалій, заснована на класифікації, розглядається при навчанні з «учителем» та при частковому навчанні з «учителем». Класифікація використовується для навчання моделі даних, віднесених до різних класів (етап навчання) та віднесення екземплярів даних до одного з наявних класів з використанням отриманої моделі (етап іспиту). Методи виявлення аномалій, засновані на класифікації, припускають, що якщо класифікатор, може бути навчений у наявному просторі ознак, то він зможе розділити нормальні та аномальні об'єкти.

До переваг методів виявлення аномалій, заснованих на класифікації, належить можливість використовувати величезну кількість методів та алгоритмів, розроблених у галузі машинного навчання – особливо для випадку, коли навчальна множина містить приклади кількох класів.

2.3 Методи класифікації за допомогою машинного навчання

2.3.1 Нейронні мережі в задачах класифікації

Нехай є m об'єктів, кожен з n параметрами. Задача k – класифікації полягає в пошуку за допомогою цих даних функції:

$$f : \mathbb{R}^n \to \{0,\ldots,K-1\},$$

що ставить будь-якому об'єкту у відповідність клас.

Нейронна мережа – це послідовність нейронів, з'єднаних між собою синапсами. Структура нейронної мережі прийшла у світ програмування прямо з біології. Завдяки такій структурі машина знаходить здатність аналізувати і навіть запам'ятовувати різну інформацію. Нейронні мережі також здатні не лише аналізувати інформацію, що входить, але й відтворювати її зі своєї пам'яті. Нейронні мережі використовуються для вирішення складних завдань, які вимагають аналітичних обчислень подібних до тих, що робить людський мозок. Найпоширенішим застосуванням нейронних мереж є класифікація [12].

Далі ми будемо розглядати нейронну мережу тільки як класифікатор. За своєю нейронна мережа суттю – це граф з n вершинами-входами і K вершинами-виходами. Тут K – кількість класів, в завданні класифікації. Вершини в цьому графі називаються нейронами, які пов'язані зваженими ребрами. Значення нейрона визначається вагами ребер, що входять в нього і значенням нейронів на протилежних кінцях цих ребер. За n параметрами об'єкту нейронна мережа видає K чисел в відрізку $[0,1]$, індекс максимального числа оголошується класом об'єкта.

Нейронна мережа прямого поширення сигналів – це штучна нейронна мережа, в якій сполучення між вузлами не утворюють циклів. На рис. 2.1 зображено нейронну мережу прямого поширення [13].

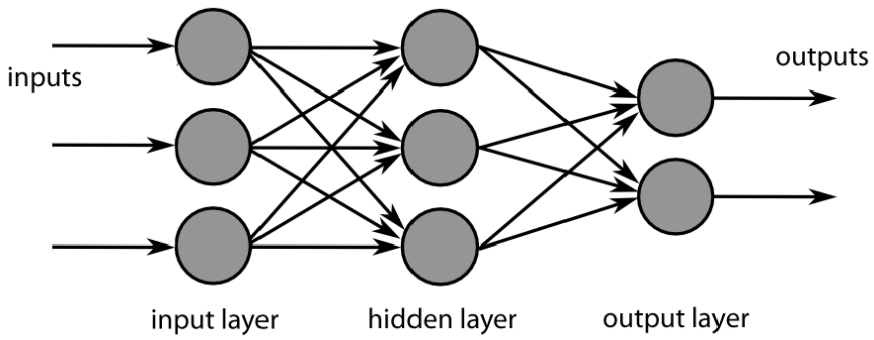


Рисунок 2.1 – Нейронна мережа прямого поширення

2.3.2 Згорткові нейронні мережі

Згорткова нейронна мережа (Convolutional Neural Network) – спеціальна архітектура штучних нейронних мереж, орієнтована на ефективне розпізнавання образів. Вона використовує деякі особливості зорової кори, у якій були відкриті прості клітини, що реагують на прямі лінії під різними кутами, та складні клітини, реакція яких пов'язана з активацією певного набору простих клітин. Таким чином, ідея згорткових нейронних мереж полягає в чергуванні згорткових шарів та субдискретизуючих шарів.

Свою назву згорткова нейронна мережа отримала за назвою операції – згортка. Згортка – це лінійне перетворення, що застосовується до квадратного вікна та плавно переміщується по вхідним даним, при цьому пікселі з вікна скалярно перемножуються на матрицю згортки.

Ідея згорткових нейронних мереж полягає в чергуванні згорткових шарів, шарів субвиборки та та шарів «звичайної» нейронної мережі. Згорткові мережі є вдалою серединою між біологічно правдоподібними мережами та звичайним багатошаровим персептроном. На сьогоднішній день найкращі результати у розпізнаванні зображень отримують за допомогою згорткових нейронних мереж. У середньому точність розпізнавання таких мереж перевищує стандартні нейронні мережі на 10-15%. Згорткові нейронні мережі – це ключова технологія Deep Learning.

Робота згорткової нейронної мережі зазвичай інтерпретується як перехід від конкретних особливостей зображення до більш абстрактних деталей, і далі до ще більш абстрактних деталей аж до виділення понять високого рівня. При цьому мережа самоналаштовується і виробляє необхідну ієрархію абстрактних ознак (послідовності карт ознак), фільтруючи незначні деталі і виділяючи суттєве.

Принцип роботи мережі: зображення проходить через ряд згорткових, нелінійних прошарків, прошарків, що об'єднують та повнозв'язних прошарків та генерується ймовірністю класів, які найкраще описують зображення. У

згортковій нейронній мережі в операції згортки використовується лише обмежена матриця ваг невеликого розміру, яку «рухають» по всьому шару, та сформовує після кожного зсуву сигнал активації для нейрона наступного шару з аналогічною позицією. Тобто для різних нейронів вихідного шару використовуються та сама матриця ваг, яку також називають ядром згортки. Тоді наступний шар, що вийшов в результаті операції згортки такою матрицею ваг, показує наявність даної ознаки в шарі, що обробляється, і її координати, формуючи так звану карту ознак.

Перший прошарок в мережі завжди згортковий. Це набір функціональних карт. Ці карти є звичаними матрицями. Кожна карта має синаптичне ядро (скануючий ядро або фільтр). Ядро є фільтром або вікном, яке рухається по всій області попередньої карти і знаходить певні ознаки об'єктів.

Нейрони на першому згортковому прошарку не зв’язані з кожним пікселем вхідного зображення, а тільки з пікселями у власних рецепторних полях. В свою чергу нейрон на другому згортковому прошарку зв’язаний тільки з нейронами, що знаходяться всередині невеликого прямокутника в першому прошарку. Така архітектура дозволяє мережі зосередитися на низькорівневих ознаках в першому прихованому прошарку, потім компонувати їх в ознаки більш високого рівня у наступному прихованому прошарку.

Коли зображення проходить через один згортковий прошарок, вихід першого рівня стає вхідним значенням 2-го рівня. Після застосування набору фільтрів будуть активовані фільтри. Вони представляють властивості вищого рівня.

Після згорткових прошарків йде прошарок, що об'єднує (прошарок понижувальної дискретизації). Основна задача прошарків, що об’єднують (pooling layer) – прорідити вхідне зображення для зменшення обчислювального навантаження, витрат пам’яті та кількість параметрів, зменшуючи ризик перенавчання. Вибір максимального прошарку (max pooling layer) є найпопулярнішим. Зазвичай, кожна карта має ядро розміром 2x2, що дозволяє зменшити попередні карти згорткового прошарку в 2 рази. Вся карта ознак

розділяється на осередки 2x2 елемента, з яких вибираються максимальні за значенням.

Останній з типів прошарку – це прошарок багатошарового персептрону. Метою є класифікація. Прошарок моделює складну нелінійну функцію, оптимізація якої підвищує якість розпізнавання. Вихідний прошарок пов'язаний з усіма нейронами попереднього прошарку. Кількість нейронів відповідає кількості розпізнаних класів [14]. На рис. 2.3 наведено приклад архітектури згорткової нейронної мережі.

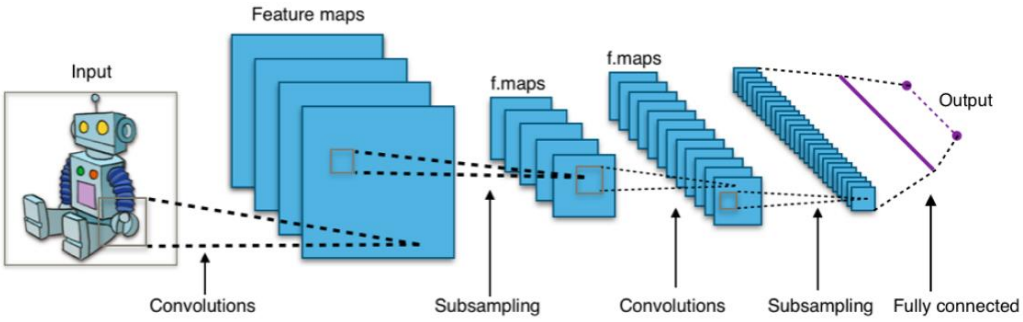


Рисунок 2.3 – Архітектура згорткової нейронної мережі

2.4 Метрики якості моделі-класифікатора

Після проведення класифікації необхідно проаналізувати результати та оцінити якість. Для цього застосовують метрики, такі як матриця невідповідностей та ROC-крива. Вони по-різному враховують відповіді класифікатора, щоб оцінити результат.

Точність (precision) та повнота (recall) є метриками, які використовуються при оцінці більшої частини алгоритмів вилучення інформації. Точність системи у межах класу – це частка об'єктів, що належать даному класу відносно всіх об'єктів, які система віднесла до цього класу. Повнота системи – це частка знайдених класифікатором об'єктів, що належать класу відносно всіх об'єктів цього класу у тестовій вибірці. Ці значення легко розрахувати виходячи з матриця невідповідностей, що складається кожного класу окремо.

У табл. 2.1 представлена матриця невідповідностей (confusion matrix).

Таблица 2.1 – Матриця невідповідностей (confusion matrix)

	$y = 1$	$y = 0$
$\hat{y} = 1$	True Positive (TP)	False Positive (FP)
$\hat{y} = 0$	False Negative (FN)	True Negative (TN)

У таблиці міститься інформація скільки разів система прийняла вірне і скільки разів помилкове рішення, де $\hat{y}$ – відповідь алгоритма класифікації на об'єкті, а y – дійсна мітка класу на цьому ж об'єкті. А саме:

– істино-позитивне рішення (TP);

– істино-негативне рішення (TN);

– хибно-позитивне рішення (FP);

– хибно-негативне рішення (FN).

Тобто помилки класифікації бувають двох типів: FP та FN.

Тоді точність та повнота розраховуються за формулами:

$$PRE = \frac{TP}{TP + FP},$$

$$REC = \frac{TP}{TP + FN}.$$

Інтуїтивно зрозумілими метриками є помилка (error, ERR) та правильність (accuracy, ACC):

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

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

$$ACC = 1 - ERR.$$

F1 – метрика, що є гармонійним середнім між точністю і повнотою, та змінюється від 0 до 1. Вона прагне до нуля, якщо точність чи повнота прагне до нуля. Обчислюється за формулою

$$F1 = 2\frac{PRE \times REC}{PRE + REC}.$$

Крива помилок або крива AUC-ROC (Area Under Curve – Receiver Operating Characteristic curve) – графічна характеристика якості бінарного класифікатора, залежність частки вірних позитивних класифікацій від частки помилкових позитивних класифікацій. Перевагою ROC-кривої є її інваріантність щодо відношення ціни помилки I та II роду. ROC-крива показує залежність TPR від FPR. ROC крива є кривою від (0,0) до (1,1) в координатах True Positive Rate (TPR) і False Positive Rate (FPR):

$$TPR = \frac{TP}{TP + FN},$$

$$FPR = \frac{FP}{FP + TN}.$$

Властивості AUC-ROC :

– в ідеальному випадку, коли класифікатор не робить помилок ($FPR = 0$, $TPR = 1$), площа під кривою дорівнює одиниці;

– якщо класифікатор видаватиме однакову кількість TP і FP, то AUC-ROC буде наближатися до 0,5;

– кожна точка на графіку відповідає вибору деякого порогу;

– площа під кривою показує якість алгоритму, тому чим більша площа, тим краще;

– крива в ідеалі повинна наближатися до точки (0,1), бо необхідно максимізувати TPR, мінімізуючи FPR.

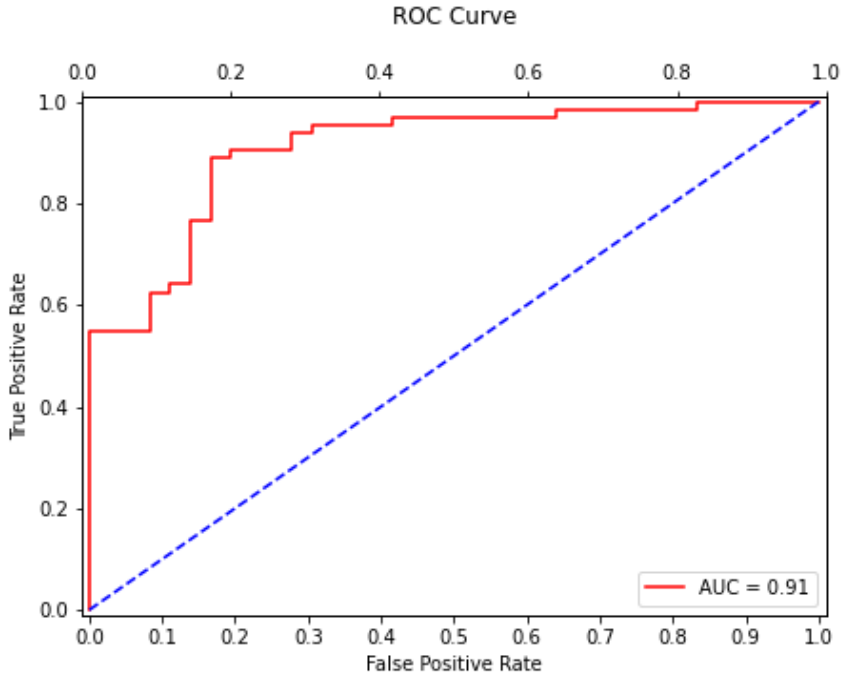


Рисунок 2.4 – Приклад кривої AUC-ROC

3 ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1 Python як високорівнева мова програмування

Python – це високорівнева об'єктно-орієнтована мовою програмування зі строгою динамічною типізацією (основна частина перевірок типів виконується під час виконання програми, а не під час компіляції), з автоматичним керуванням пам'яттю, орієнтований на підвищення продуктивності розробника, читання коду та його якості, і навіть забезпечення переносимості написаних у ньому програм. Python є мовою з відкритим кодом, та містить багато бібліотек для обробки та графічного зображення даних. Він може підтримувати основні структури даних: кортежі, списки, масиви, словники.

Однією з найважливіших переваг є те, що Python працює на всіх відомих платформах, таких як Windows, UNIX, MAC OS, MAC OS X, IPhone OS, Android. Мова є повністю об'єктно-орієнтованою – все є об'єктами. Незвичайною особливістю мови є виділення блоків коду пробільними відступами. Синтаксис ядра мови мінімалістичний, тому рідко виникає необхідність звертатися до документації.

Також Python є швидко розширювальною мовою програмування. Тобто мова має можливість вдосконалення внаслідок роботи усіма програмістами, які прагнуть вдосконалити її. У 2008 році з’явилося велике оновлення мови. Раніше був Python 2, а після цього розробили новий Python 3. Він є більш сучасним і майбутнім мови. Нова версія Python 3 зберегла повну сумісність з більш старими варіантами, але у новій версії були усунено багато ключових недоробок в архітектурі. Всі останні покращення стандартної бібліотеки доступні за замовчуванням лиши в Python 3.

Стандартна бібліотека включає великий набір корисних функцій, починаючи від функціоналу для роботи з текстом та закінчуючи засобами для написання мережевих додатків. Додаткові можливості, такі як математичне моделювання, робота з обладнанням, написання веб-додатків або розробка ігор

можуть бути реалізовані за допомогою великої кількості сторонніх бібліотек. Також перевагою є наявність великого числа бібліотек, що забезпечують різні додаткові можливості. Такі бібліотеки можуть бути написані усіма кваліфікованими програмістами. Як приклад можна навести модуль Numerical Python, якій має розширені математичні можливості, такі як робота з цілими векторами та матрицями.

Python має дуже докладну документацію. При виникненні питання необхідно скористатися стандартною бібліотекою довідки pydoc. Для доступу до неї досить викликати функцію help, після якої в дужках як аргумент вказати, що саме цікавить. Також можна скористатися можливостями pydoc для створення власної документації, наприклад якщо потрібно задокументувати новий модуль. Недоліками мови є найчастіше нижча швидкість роботи та більш високе споживання пам'яті написаних на ньому програм.

Python відноситься до найбільш затребуваною і популярною мовою програмування, про що свідчать численні рейтинги та аналіз пропозицій на ринку розробки програмних продуктів.

3.2 Алгоритм розв’язання задачі класифікації та виявлення аномалій часових рядів

Для розв’язання задачі класифікації медичних часових рядів, з метою пошуку аномалій, було виконано експеримент на основі побудови рекурентних діаграм. Алгоритм виконання експерименту:

– пошук датасету медичних часових рядів на основі ЕКГ;

– завантаження датасету з рядами ЕКГ, кожна серія якого простежує електричну активність, записану під час серцебиття;

– серіалізація даних;

– побудова зображення рукурентної діаграми для кожного ряду;

– побудова багатошарової нейронної мережі для розпізнання зображень

рекурентних діаграм;

– компіляція моделі нейронної мережі;

– проведення навчання згорткової нейронної мережі;

– оцінка якості навчання нейронної мережі на тестових даних з датасету;

– побудова ROC-кривої;

– проведення аналізу класифікації медичних часових рядів;

– формулювання висновку про проведену роботу на основі отриманих даних.

3.3 Опис програми

Програма написана високорівневою мовою програмування Python в середовищі Colaboratory, яке використовує Google Cloud. Вона класифікує рекурентні діаграми, отримані з часових рядів ЕКГ, з метою виявлення аномалій. Мета програми – отримати класифікатор, який говорив би нам, з деякою часткою ймовірності, де нормальні ряди ЕКГ, а де знайдено аномалію.

Для тренування класифікатора маємо тільки набір даних з «нормальними» рядами ЕКГ. У першій частині роботи ми розробили клас, який за візуалізує часовий ряд, будує рекурентні діаграми, зберігає графіки у файли. Також він проводить необхідні розрахунки та зберігає рекурентну діаграму у вигляді numpy масиву, який можна застосовувати як вхід для нейронної мережі.

На рис. 3.1 наведено приклади методів класу для відображення рекурентних діаграм та зберігання зображень рекурентних діаграм.

Далі завантажили у зручному вигляді датасет та зробили серіалізацію. Зробили зображення рекурентної діаграми для кожного ряду. Приклад часового ряду ЕКГ та його рекурентної діаграми наведено на рис. 3.2.

Далі додали багатошарову згорткову нейронну мережу застосовуючи бібліотеку «Keras», яка є найпопулярнішою для створення нейронних мереж. Приклад створення згорткової нейронної мережі наведено на рис. 3.3.

```
def reccurence_plot(self):
  plt.imshow(self.recurrence_matrix_inv , cmap='gray', origin='lower')
  plt.show()


def save_image_rec(self, filename):
  imageio.imwrite(filename, (255*self.recurrence_matrix_inv).astype(np.uint8))
```

Рисунок 3.1 – Методи класу для відображення та зберігання діаграм

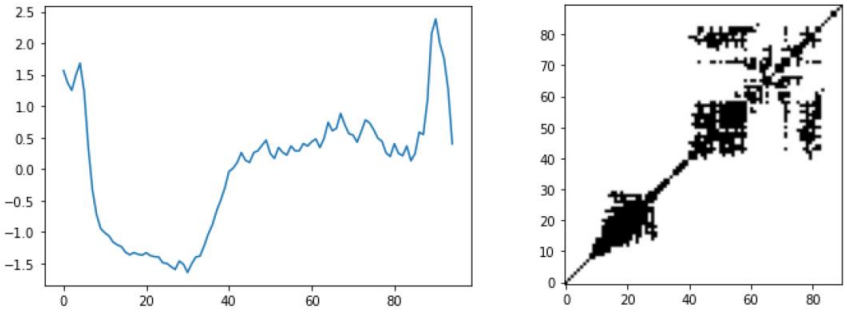


Рисунок 3.2 – Часовий ряд та рекурентна діаграма отримана з цього ряду

```
model = Sequential()

model.add(Conv2D(32, kernel_size = (5,5),
                 activation ='relu', input_shape = input_shape))
model.add(Conv2D(32, kernel_size = (5,5),
                 activation ='relu'))
model.add(MaxPooling2D(pool_size=(2,2)))
model.add(Dropout(0.25))


model.add(Conv2D(64, kernel_size = (3,3),
                 activation ='relu'))
model.add(Conv2D(64, kernel_size = (3,3),
                 activation ='relu'))
model.add(MaxPooling2D(pool_size=(2,2)))
model.add(Dropout(0.25))

model.add(Flatten())
model.add(Dense(128, activation = 'relu'))
model.add(Dropout(0.5))
model.add(Dense(1, activation = 'sigmoid'))

model.compile(loss='binary_crossentropy',
              optimizer='Adadelta',
              metrics=['accuracy'])
```

Рисунок 3.3 – Згорткова нейронна мережа у Python

Для визначення точності класифікаторів ми використовували різні метрики, такі як accuracy, presicion, recall, f1-score, AUC. Наводили графіки точності під час навчання та ROC-криві.

4 РЕЗУЛЬТАТИ ОБЧИСЛЮВАЛЬНОГО ЕКСПЕРИМЕНТУ

4.1 Постановка задачі

Метою експерименту є виявлення аномальних імпульсів серцевого ритму з використанням бази даних ЕКГ. Дані ЕКГ можна розглядати як періодичні часові ряди, аномалія в цьому випадку була б невідповідною картиною, наприклад, з точки зору періодичності або амплітуди, що може свідчити про проблему зі здоров'ям.

Для обчислювального експерименту обрали датасет «ECG200» з часовими рядами для класифікації з репозиторію «UEA & UCR Time Series Classification Repository».

В датасеті «ECG200» містяться медичні часові ряди, що отримані з електрокардіограми. Ряди розбиті на два класи: «норма» та «аномалії», рис. 4.1. В датасеті є 200 семплів: 100 для тренування класифікатора, у якому тільки «нормальні» ряди, та 100 для перевірки з «нормальними» рядами та з «аномаліями». В кожному семплі по 100 значень.

Цей набір даних був відформатований Р. Ольшшевським у рамках дипломної роботи "Узагальнене вилучення особливостей для розпізнавання структурних зразків у даних часових рядів" у 2001 році. Кожна серія простежує електричну активність, записану під час серцебиття.

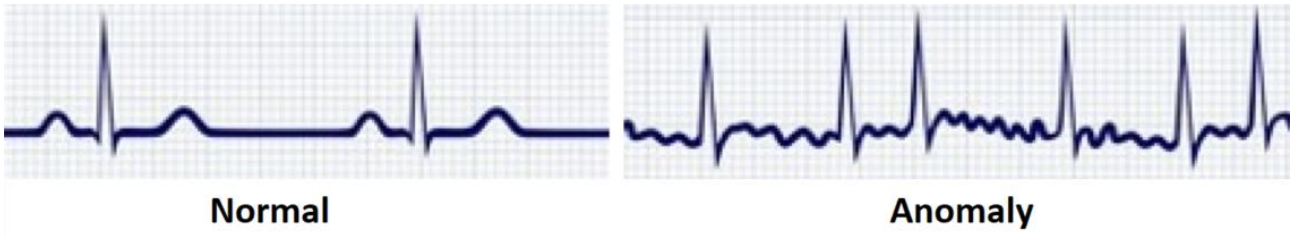


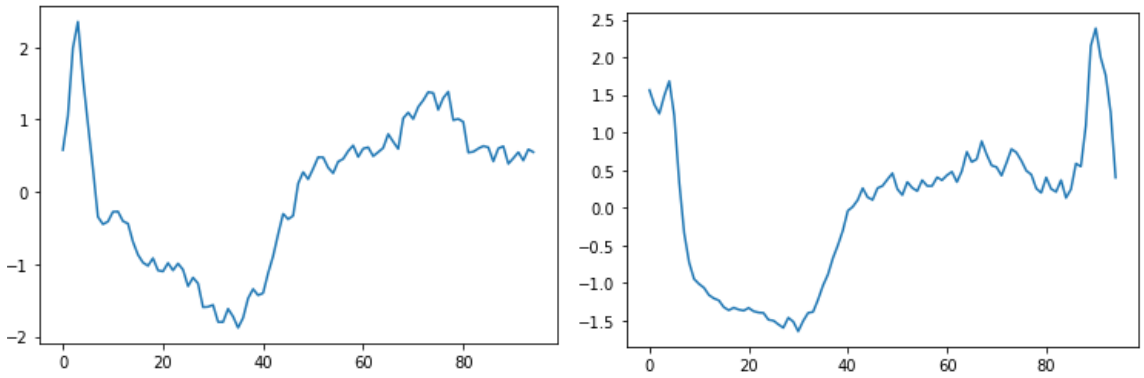
Рисунок 4.1 – ЕКГ, «норма» та «аномалія»

Оскільки у тренувальному наборі ми маємо інформацію тільки про нормальні часові ряди, а приклади аномалій відсутні, то можна вирішити таку

задачу як задача класифікації з нормальною поведінкою серцебиття аномальною. Тобто потрібно визначити, чи є часовий ряд тестувальної вибірки аномальним по відношенню до бази даних часових рядів, який складається тільки з нормальних часових рядів.

Для експерименту застосуємо багатошарову згорткову нейронну мережу, оскільки класифікація буде проводитись на зображеннях рекурентних діаграм.

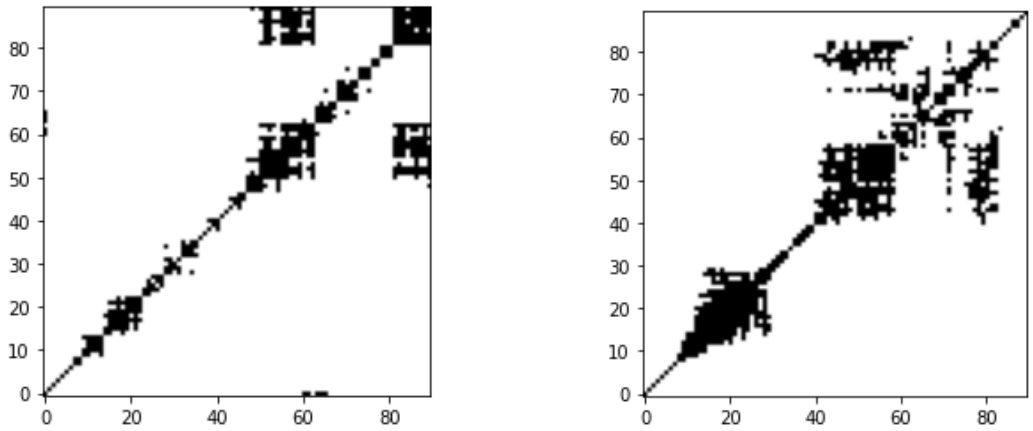
На рис. 4.2 наведено приклади часових рядів з тестової вибірки, а на рис. 4.3 наведено приклади відповідних рекурентних діаграм.



а) б)

а) «нормальні» ряди ЕКГ; б) «аномальні» ряди ЕКГ

Рисунок 4.2 – Приклади часових рядів з тестового датасету



а) б)

а) «нормальні» ряди; б) «аномальні» ряди

Рисунок 4.3 – Рекурентні діаграми часових рядів

У табл. 4.1 наведено кількості зразків кожного класу.

Таблиця 4.1 – Розподіл зразків для класифікації

Dataset	Normal	Anomaly	Total
Train	100	0	100
Test	36	64	100
Total	136	64	200

4.2 Проведення експерименту

Натренуємо класифікатор за допомогою «рекурентних діаграм». Для побудови рекурентних діаграм оберемо такі значення параметрів:

$$\varepsilon = 0,25,\ \tau = 5.$$

На рис. 4.4 наведено метод для відображення рекурентних діаграм та зберігання зображень рекурентних діаграм.

```
# saving reccurence plots (test)
# adding images to dataset (test)

images_test = []
rec_test = []

for i, series in enumerate(test_series):
  rec = RecurrenceTimeSeries(series, tau, radius, 2, 2)
  # зберігаємо локально зображення у вигляді файлів
  filename_1 = 'rec_plots/test/' + ('0' if test_labels[i] == 0.0 else '1') + '/' + str(i) + '_rec' + '.png'
  filename_2 = 'rec_plots/test/' + ('0' if test_labels[i] == 0.0 else '1') + '/' + str(i) + '_ser' + '.png'
  rec.save_image_rec(filename_1)
  rec.save_image_series(filename_2)
  # додаємо в список
  images_test.append(rec.recurrence_matrix)
  rec_test.append(rec)

images_test = np.array(images_test)
```

Рисунок 4.4 – Метод для відображення рекурентних діаграм з часових рядів

Приклад отриманих рекурентних діаграм наведено на рис. 4.5 та 4.6. Розглядаючи графіки, можна помітити візуальні відмінності. Для «нормальних» часових рядів на діаграмі праворуч знизу квадрат, для «аномальних» – дещо інший вигляд.

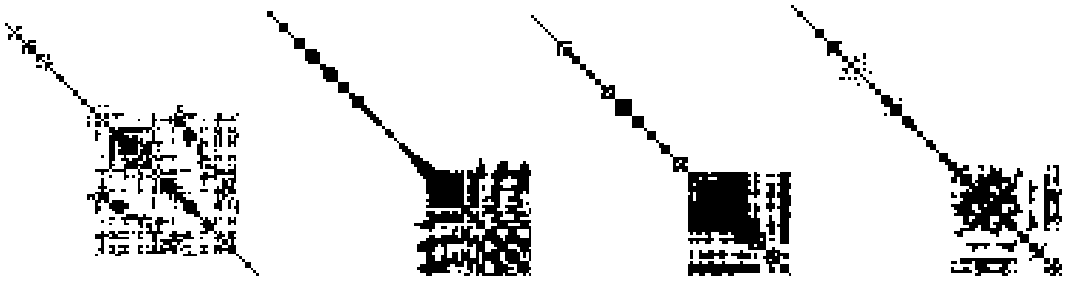


Рисунок 4.5 – Рекурентні діаграми для «нормальних» рядів

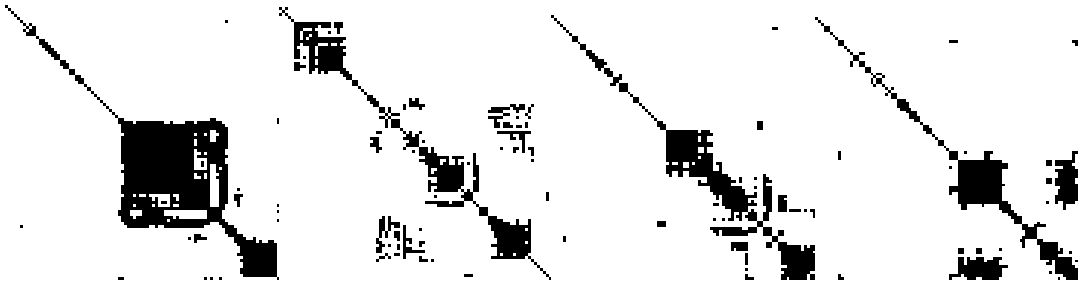


Рисунок 4.6 – Рекурентні діаграми для «аномальних» рядів

У якості класифікатора візьмемо багатомасштабну згорткову нейронну мережу. На вхід нейронна мережа приймає зображення (двомірний масив чисел) рекурентної діаграми часового ряду, який має бути передбачений. А вихід – це його мітка, тобто число у діапазоні [0,1]. Основна відмінність згорткової нейронної мережі від персиптрону полягає у згорткових прошарках, які позначені як «Conv2D» на схемі.

Структура згорткової нейронної мережі має три послідовні етапи: трансформація, локальна згортка та повномасштабна згортка.

Вхідні дані являють собою зображення рекурентних діаграм розміру 90 на 90 пікселів. Візуалізація згорткової нейронної мережі зображена на рис. 4.7.

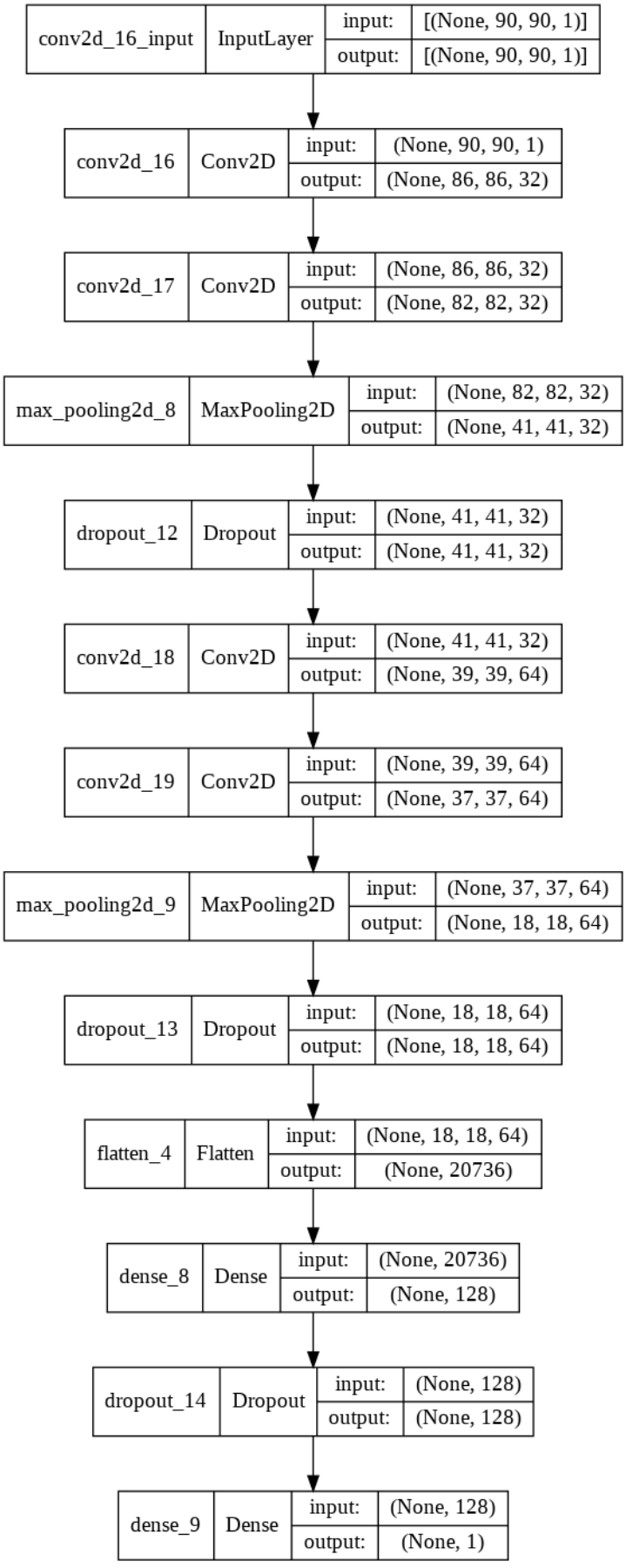


Рисунок 4.7 – Схема згорткової нейронної мережі

На етапі трансформації застосовуються різні перетворення у вхідних часових рядах. Проводиться перетворення прямого відображення, перетворення зі зменшенням вибірки у часовій області та спектральні перетворення в частотній області. Кожна частина називається гілкою, так як вони є гілками, що входять у згорткову нейронну мережу.

На етапі локальної згортки використовується кілька шарів для вилучення компонентів для кожної галузі. На цьому етапі згортки для різних гілок незалежні один від одного. Усі виходи проходять через процедуру субдескритизацію (max pooling) із кількома розмірами.

На стадії повномасштабної згортки поєднуються всі вилучені ознаки та застосовується ще кілька згорткових шарів (кожен з яких слідує за субдескритизацією), повнозв'язні шари та шар softmax використовуються для створення кінцевого результату. Усі параметри навчаються спільно за допомогою зворотного розповсюдження помилки.

4.3 Точність роботи класифікатора

Для тестування алгоритмів класифікації на вхід класифікатора подавалися різні часові ряди, що описували нормальну роботу серця. Класифікатор, що було побудовано на багатошаровій згортковій нейромережі, тренували 120 епох. Точність перевірили на тестовій вибірці, яка містить як «нормальні» так і «аномальні» часові ряди коливань серця. Точність на тестовій вибірці сягнула близько 75% та більше не збільшувалась. Точність різними метриками наведена у табл. 4.2.

На рис. 4.8 візуалізували якість моделі упродовж навчання. Як бачимо, точність на тренуваній виборці трохи більша, ніж на тестовій. Це може свідчити про статистичну відмінність між вибірками, бо у навчальній виборці містяться «нормальні» медичні часові ряди, також розбиття на «train» та «test» було виконано одразу в датасеті.

Таблиця 4.2 – Точність класифікатора різними метриками

	Precision	Recall	F1-score	Support
Normal	0,80	0,44	0,57	36
Anomaly	0,75	0,94	0,83	64
Accuracy			0,76	100
Macro avg	0,78	0,69	0,70	100
Weighted avg	0,77	0,76	0,74	100

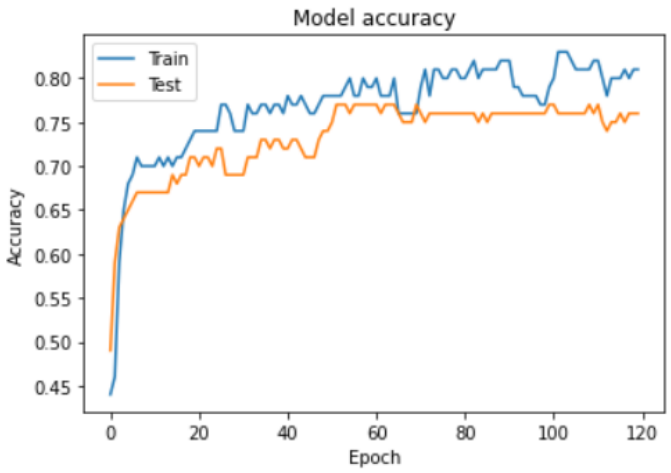


Рисунок 4.8 – Рисунок точність моделі під час навчання

Метрика AUC дорівнює 0,76, графік ROC кривої наведено на рис. 4.9. Відомо, що чим вище червона лінія знаходиться від середини (синя пунктирна лінія) тим якісніша класифікація (ідеальний варіант коли крива майже збігається з прямою $y \approx 1$).

Оскільки червона лінія знаходиться вище за синю пунктирну лінію, отримали задовільний результат класифікації. Судячи з двох графіків точності моделі упродовж навчання, точність на тренувальних даних трохи вища ніж на тестових. Так вібдувається через те, що в данному експерименті проводиться навчання частково з «учителем».

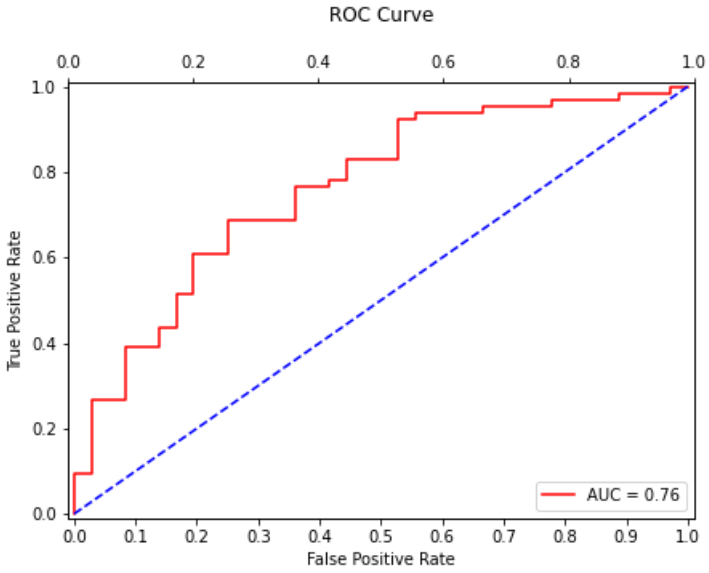


Рисунок 4.9 – ROC крива

Якщо у тренувальній виборці подавати інформацію як про «нормальну» так і про «аномальну» поведінку, точність була вищою. Але, зазвичай, у медичних центрах актуальним є пошук саме усіх аномальних поведінок часового ряду ЕКГ, аби відстежити будь-які відхилення у ритмі серця.

ВИСНОВКИ

Виявлення аномалій серцевих ритмів на ранніх стадіях захворювання може врятувати життя людині. Виконання даної роботи є актуальним для лікувальних установ, бо класифікація ЕКГ дозволяє виявити взаємозв'язок між її характеристиками та видом захворювання, та запропонувати найбільш ефективний спосіб діагностування.

Існує багато методів для пошуку аномалій часових рядів. У даній роботі розглядається метод пошуку аномалій, заснований на класифікації. Для класифікації було застосовано методи машинного навчання, що у наш час є актуальним. За допомогою машинного навчання можна дуже швидко класифікувати ряди ЕКГ. Машинне навчання – це найсучасніший метод метод класифікації часових рядів. Класифікатор, що отримали у наслідок виконання роботи, можна використовувати для виявлення хвороб у пацієнта. Він виявляє чи є у часовому ряді аномалія, чи він у нормі. Але дана робота не є кінцевим продуктом, який можуть використати у медичних закладах, бо програма потребує доробки.

Цей спосіб виявлення аномалій ЕКГ полегшить роботу лікарів під час діагностування хвороб серця. Саме тому, дана робота є затребуваною у галузі медицини.

В ході даної кваліфікаційної роботи було досліджено метод виявлення аномалій та класифікації медичних часових рядів, заснований на побудові рекурентних діаграм з використанням багатомасштабної згорткової нейронної мережі.

Було розглянуто задачу класифікації медичних часових рядів, використовуючи датасет «ECG200», в якому містяться медичні часові ряди, що отримані з електрокардіограми. Ряди розбиті на два класи: «норма» та «аномалії». В якості класифікатора була багатомасштабна згорткова нейронна мережа.

Були зібрані тренувальні дані, які містили інформацію тільки про

нормальну поведінку серцебиття людини, які необхідні для нейронної мережі. Також була отримана тестова вибірка для її подальшої класифікації. Також було розроблено ряд алгоритмів, спрямованих на обробку даних, виділення ознак, притаманних певним компонентам, та нормуванням їх для подальшої класифікації нейронною мережею.

Як показали результати, розглянутий метод пошуку аномалій має досить високу точністю класифікації. Точність класифікатора із застосуванням зображень рекурентних діаграм сягнула 75%. Результати роботи можуть бути використані для виявлення аномалій та для класифікації часових рядів по методам машинного навчання.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Степаненко Ю. С. Визначення аномалій у часових рядах на основі візуалізації рекурентних діаграм // 25-й Міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті» : зб. матеріалів форуму (м. Харків, 20-22 квітня 2021 р.). Т. 7. Харків : ХНУРЕ, 2021. С. 69–70.

2. Кириченко Л. О., Степаненко Ю. С., Яндуков Д. А. Класифікація часових рядів із використанням рекурентних діаграм // Системні технології. 2021. № 136. С. 81–87.

3. Applying recurrence plot to classify time series. URL : http://ceur-ws.org/Vol-2870/paper128 (дата звернення: 07.10.2021).

4. Джефферс Д. Введение в системный анализ: применение в экологии. Москва : Мир, 1981. 252 с.

5. Колмогоров А. Н., Фомин С. В. Элементы теории функций и функционального анализа. Москва : Наука, 1968. 543 с.

6. Прикладная статистика: Классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. Москва : Финансы и статистика, 1989. 607 с.

7. Recurrence plots for the analysis of complex system / N. Marwan, M. Romano, M. Thiel, J. Kurths // Physics Reports. 2007. № 1. P. 237–329.

8. Kirichenko L. O., Kobitskaya Y. A., Habacheva A. Y. Comparative Analysis of the Complexity of Chaotic and Stochastic Time Series // Radioelectronics. Informatics. Management. 2014. № 2. P. 126–134.

9. Кириченко Л. О., Кобицкая Ю. А., Хабачева А. Ю. Сравнительный рекуррентный анализ хаотических и случайных процессов // Physical and technological problems of radio engineering devices, telecommunication, nano- and microelectronics. Proceeding of the II International Scientific-Practical Conference. 2013. № 2. P. 48–52.

10. Kirichenko L., Zinchenko P., Radivilova T. Machine Learning Detection of DDoS Attacks Based on Visualization of Recurrence Plots // Radioelectronics. In-

formatics. 2019. № 2. P. 23–34.

11. Time series classification. URL : http://www.timeseriesclassification.com (дата звертання: 23.10.2021).

12. Машинное обучение (курс лекций, К.В.Воронцов). URL : http://www.machinelearning.ru (дата звертання: 27.10.2021).

13. Методы классификации и прогнозирования. URL : http://www.intuit.ru /studies /courses/6/6/lecture/182 (дата звертання: 05.11.2021).

14. Краткий обзор языка Python. URL : https://www.helloworld.ru /texts/comp/lang/python (дата звертання: 15.11.2021).