ISSN 2519-2310 (Online)

КОМП'ЮТЕРНІ НАУКИ ТА КІБЕРБЕЗПЕКА, 2024, № 1 (25)

COMPUTER SCIENCE AND CYBERSECURITY, 2024, Iss. 1 (25)

DOI: https://doi.org/10.26565/2519-2310-2024-1-04
УДК 004.056.5

КЛАСТЕРИЗАЦІЯ ТА КЛАСИФІКАЦІЯ ЧАСОВИХ ЗВУКОВИХ РЯДІВ

Станіслав Качанов[1], аспірант, e-mail: staskachanov2000@gmail.com,
ORCID: https://orcid.org/0009-0002-6938-6717
Дмитро Власенко[1], старший викладач кафедри теоретичних та прикладних комп'ютерних наук, кандидат математичних наук, e-mail: vlasenkod@karazin.ua,
ORCID: https://orcid.org/0009-0006-8780-2066
[1]*Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 4, Харків, 61022, Україна*

Рукопис надійшов 17 березня 2024 р. Отримано після рецензування 19 квітня 2024 р.
Прийнято 20 травня 2024 р.

Анотація: Було розглянуто дві важливі задачі в аналізі даних – класифікація та кластеризація часових рядів на прикладі звукових записів серцебиття людей. Однією з основних проблем аналізу часових рядів є складність порівняння різних рядів через їх варіативність у довжині, формі та амплітуді коливань. Для вирішення цих задач були використані різні алгоритми, серед яких рекурентна нейронна мережа з довгою короткочасною пам'яттю (LSTM) і алгоритм k найближчих сусідів для класифікації, та метод k-середніх (K-means) і DBSCAN для кластеризації. Результати дослідження показали, що LSTM є потужним інструментом для класифікації часових рядів завдяки здатності зберігати інформацію про контекст у часі. KNN, з іншого боку, продемонстрував високу точність і швидкість класифікації, однак його обмеження проявилися в умовах великих наборів даних. Для задач кластеризації, метод K-means виявився більш ефективним у порівнянні з DBSCAN, демонструючи вищу якість кластеризації за метриками силуету, Rand Score та іншими. Дані для дослідження були отримані з архіву часових рядів UCR, що включає звукові записи серцебиття різних категорій. Аналіз результатів показав, що обрані методи класифікації та кластеризації можуть бути ефективно використані для діагностики серцевих захворювань. Крім того, це дослідження відкрило нові можливості для подальшого вдосконалення методів обробки та аналізу даних, зокрема, для розробки нових інструментів медичної діагностики. Таким чином, ця робота демонструє ефективність використання алгоритмів машинного навчання для аналізу часових рядів та їх значення для покращення діагностики серцево-судинних захворювань.

Ключові слова: *класифікація часових рядів, кластеризація часових рядів, рекурентна нейронна мережа, LSTM, KNN, K-means, DBSCAN, аналіз звукових даних, серцеві звуки, машинне навчання, діагностика серцевих захворювань*

Як цитувати: Качанов С., Власенко Д.. Кластеризація та класифікація часових звукових рядів. *Комп'ютерні науки та кібербезпека.* 2024; № 1(25): С. 42–52. https://doi.org/10.26565/2519-2310-2024-1-04

In cites: Kachanov S., Vlasenko D. (2024). Clustering and Classification of Time Series Sound Data. *Computer Science and Cybersecurity.* 1(25): 42–52. https://doi.org/10.26565/2519-2310-2024-1-04 (in Ukrainian)

1. Вступ

У статті буде досліджуються дві актуальні задачі: класифікація і кластеризація часових рядів, які є важливими завданнями в аналізі даних.

Класифікація та кластеризація часових рядів є складними завданнями, оскільки ряди можуть мати різну довжину, форму та амплітуду коливань, що робить непростим встановлення схожості між рядами. У зв'язку з цим, розроблено багато різних методів для вирішення цих завдань, включаючи методи на основі статистичних моделей, нейронних мереж та інші. Існує багато алгоритмів, призначених для класифікації часових рядів. В роботі досліджено основні з них, та обгрунтовано, чому саме ці алгоритми, моделі та методи є найкращими для вирішення цих задач.

2. Опис обраних алгоритмів та підходів для задач класифікації

У цьому дослідженні як основний підхід застосовується один із найефективніших методів для розв'язання поставленої задачі за різними оцінками — використання глибокого навчання, зокрема рекурентної нейронної мережі з довготривалою короткочасною пам'яттю (LSTM) [1].

LSTM є одним із найвідоміших типів рекурентних нейронних мереж (RNN) і використовується для обробки послідовних даних, таких як мовлення, текст та часові ряди. Головна проблема традиційних RNN полягає в їхній неспроможності зберігати інформацію на довгих часових відрізках у послідовності [10]. LSTM була розроблена для вирішення цієї проблеми; вона має спеціальну структуру, яка дозволяє зберігати та використовувати інформацію протягом тривалих періодів часу. У звичайних нейронних мережах для класифікації зазвичай використовується пряме поширення сигналу (feedforward), де кожен вхідний сигнал обробляється окремо. Такі моделі не враховують динаміку змін у часі, оскільки не мають пам'яті про попередні вхідні дані.

Натомість LSTM здатна зберігати та використовувати попередні стани внутрішніх блоків, що називається пам'яттю LSTM. Це дозволяє моделі зберігати інформацію про попередній контекст і розуміти часові залежності між даними. Таким чином, LSTM може робити передбачення на основі повної історії часового ряду, враховуючи всі попередні значення, що робить її ефективною в задачах класифікації часових рядів. Крім того, LSTM може автоматично визначати, яку інформацію слід забути, а яку зберегти в пам'яті, що дозволяє моделі відкинути зайві дані, які можуть заважати правильній класифікації.

Для класифікації часових рядів за допомогою LSTM необхідно спочатку створити модель LSTM з відповідними вхідними та вихідними шарами [2]. Вхідний шар повинен мати розмірність (кількість часових кроків, кількість ознак), де кількість кроків відповідає довжині часового ряду, а кількість ознак — кількості характеристик у кожному кроці часу. У цьому дослідженні використовується набір даних із 5 класами, тому вихідний шар представляє ймовірності належності до певного класу [10].

Для порівняння також обрано алгоритм k-найближчих сусідів (k-Nearest Neighbors, KNN) — метод машинного навчання без учителя, який використовується для класифікації та регресії. Для класифікації нового часового ряду KNN знаходить k найближчих часових рядів із навчального набору, використовуючи відстань між векторами, яку можна обчислити за допомогою різних метрик, таких як евклідова або манхеттенська відстані. Класифікація нового часового ряду здійснюється шляхом голосування найближчих k сусідів. Однією з головних переваг KNN у класифікації часових рядів є його простота та ефективність [3]. Він не вимагає великої кількості навчальних даних і може добре працювати із зашумленими даними. Крім того, KNN може ефективно працювати з великими наборами даних, оскільки не потребує часу на

тренування моделі.

Однак KNN має й деякі недоліки [4]. Він може бути чутливим до викидів у даних, оскільки не враховує структуру даних і може помилково класифікувати тестові зразки, якщо навчальний набір містить зміщені дані або викиди. Щодо застосування KNN для класифікації часових рядів, цей метод може бути особливо ефективним, оскільки допомагає знаходити схожість між часовими рядами та використовувати цю інформацію для класифікації нових даних.

Таким чином, вибір методу для класифікації часових рядів повинен базуватися на конкретній задачі та характеристиках даних, які потрібно аналізувати.

3. Опис обраних алгоритмів та підходів для задач кластеризації

Метод к-середніх (K-means) є одним з найпоширеніших методів кластеризації [5], який дозволяє розділити множину даних на кластери на основі схожості між їх елементами.

Використання методу к-середніх для кластеризації часових рядів полягає в тому, щоб розділити множину часових рядів на кластери на основі їх схожості. Для цього, зазвичай, використовують такі метрики, як Евклідова відстань або манхеттенська відстань, або косинусна відстань між векторами. Крім того, можна використовувати різні методи для побудови векторів-ознак для часових рядів, такі як метод головних компонент.

У задачах класифікації часових рядів метод к-середніх можна використовувати для попередньої кластеризації даних та подальшого застосування методів класифікації для кожного кластеру окремо. Це може допомогти поліпшити якість класифікації та знизити час роботи алгоритму [6].

Для порівняння було обрано інший алгоритм кластеризації - DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

Одна з головних переваг DBSCAN полягає в тому, що він може працювати з даними різної густини та форми кластерів, і він добре підходить для кластеризації часових рядів, де зазвичай зустрічаються складні форми та різні рівні густини даних. Крім того, DBSCAN може ідентифікувати шумові точки, які не належать до жодного кластера.

Однак, існує деяка кількість недоліків, пов'язаних з використанням DBSCAN. Наприклад, якщо розмір набору даних дуже великий, то алгоритм може стати надто повільним (обраний набір даних не є дуже об’ємним, тому алгоритм виконується досить швидко).

У порівнянні з методом k-середніх, DBSCAN має деякі переваги. Наприклад, DBSCAN може працювати з даними з різною густотою та формою, тоді як k-середніх передбачає, що кластери мають сферичну форму та рівномірну густину [7].

Можна заявити з упевненістю, що DBSCAN - це потужний метод кластеризації часових рядів, який дозволяє виявляти кластери будь-якої форми та не вимагає заздалегідь визначених кількості кластерів. Його недоліки полягають у потребі в налаштуванні гіперпараметрів та високих обчислювальних витратах порівняно з методом k-середніх, але варто зазначити, що і в методі k-середніх підбір гіперпараметрів теж відіграє важливу роль.

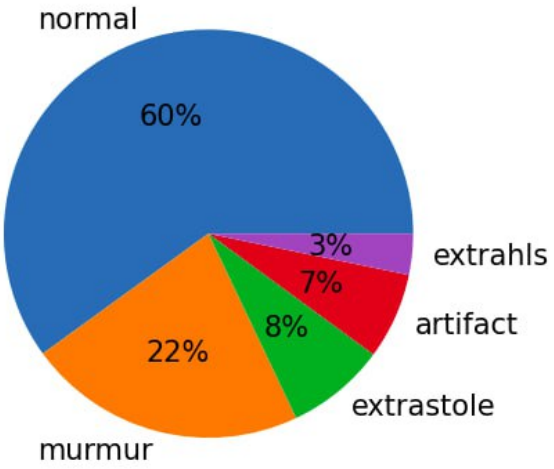
4. Актуальність обрання даних та вирішення цієї задачі, опис даних

Для обрання максимально релевантних даних, було вирішено обрати відомий архів, в якому зібрані дані за останні 20 років (синтетичні та натуральні), які найкраще підходять для наукових досліджень. Архів часових рядів UCR [8], запроваджений у 2002 році, став важливим ресурсом для спільноти з видобутку даних часових рядів.

Аудіофайли мають різну довжину від 1 до 30 секунд (деякі були обрізані, щоб зменшити надмірний шум та надати виокремлений фрагмент звуку). Більшість інформації у звуках серця

міститься у низькочастотних компонентах, а шум - у вищих частотах. Кожен файл відноситься до однієї із категорій даних:

- **Нормальний звук (normal).** У категорії «Нормальний звук» є нормальні, здорові звуки серця. Вони можуть містити шум у останній секунді запису, коли прилад віднімається від тіла. Вони можуть містити різноманітні фонові шуми (від транспорту до радіо). Вони також можуть містити випадковий шум, що відповідає диханню або торканню мікрофона одягом або шкірою.
- **Категорія бурчання (murmur).** Серцеві шуми звучать так, ніби є «свист, рев, гуркіт або бурхлива рідина» в одному з двох тимчасових місць: (1) між «луб» і «даб» або (2) між «даб» і «луб». Вони можуть бути симптомом багатьох захворювань серця, деякі серйозні.
- **Категорія додаткового серцевого звуку (extrahls).** Додатковий серцевий звук може не бути ознакою хвороби. Однак у деяких ситуаціях це важливий ознака хвороби, яку, якщо виявити рано, може допомогти людині. Додатковий серцевий звук важливо виявляти, оскільки його не можна добре виявити ультразвуком.
- **Категорія «Артефакт» (artifact).** В цій категорії є широкий спектр різних звуків, включаючи звук зворотнього зв'язку і ехо, мову, музику та шум. Зазвичай немає відчутних звуків серця і має мало або жодної тимчасової періодичності на частотах нижче 195 Гц. Ця категорія найбільш відрізняється від інших. Відрізняти цю категорію від чотирьох інших, дуже важливо щоб той, хто збирає дані, здійснив повторну спробу.
- **Категорія «Надзвичайний серцевий ритм» (extrastole).** Звуки цього ритму можуть з'являтися час від часу і можуть бути визначені тим, що серцевий ритм порушений через додаткові або пропущені серцеві скорочення (це не те саме, що додатковий серцевий звук, оскільки ця подія не відбувається регулярно). Надзвичайний серцевий ритм може не бути ознакою хвороби, однак у деяких ситуаціях надзвичайні ритми можуть бути спричинені серцевими захворюваннями.



Діаграма 1 – Відсотковий розподіл даних за категоріями
Diagram 1 – Percentage Distribution of Data by Categories

На круговій діаграмі вище (Діаграма 1) видно розподіл даних за категоріями: найбільше нормальних звичайних записів серцебиття (60%), далі йде категорія «бурчання» (22%), за нею записи з надзвичайним серцевим ритмом (8%), далі некоректні записи-артефакти (7%) і найменше з категорії додаткового серцевого звуку.

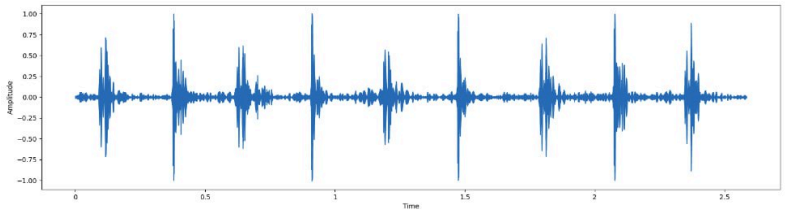


Рис. 1 – Звукова хвиля для нормального серцебиття (normal)
Fig. 1 – Sound Wave for Normal Heartbeat (normal)

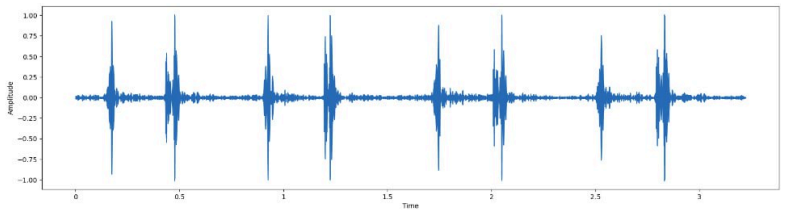


Рис. 2 – Звукова хвиля для серцебиття з бурчанням (murmur)
Fig. 2 – Sound Wave for Heartbeat with Murmur (murmur)

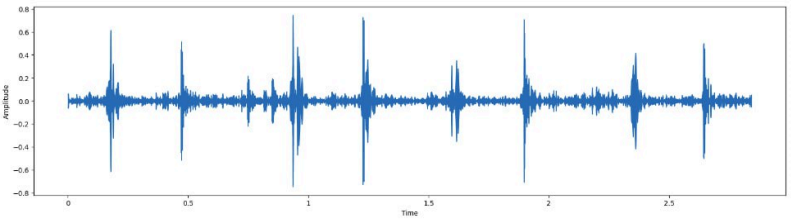


Рис. 3 – Звукова хвиля для серцебиття з надзвичайним серцевим ритмом (extrastole)
Fig. 3 – Sound Wave for Heartbeat with Extrastole (extrastole)

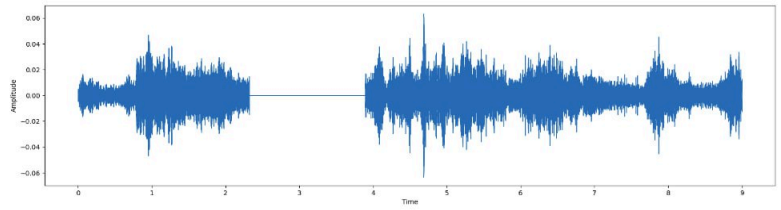


Рис. 4 – Звукова хвиля для неправильних записів (artifacts)
Fig. 4 – Sound Wave for Incorrect Recordings (artifacts)

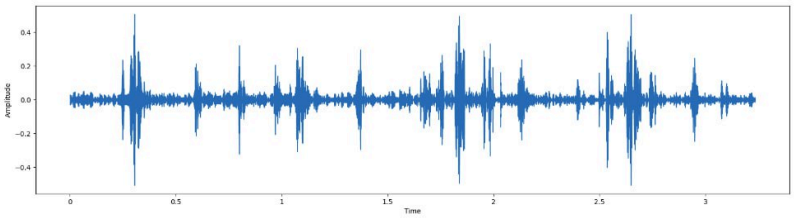


Рис. 5 – Звукова хвиля для серцебиття з додатковим серцевим звуком (extrahls)
Fig. 5 – Sound Wave for Heartbeat with Additional Heart Sound (extrahls)

На Рис. 1 добре видно яскравий цикл при нормальному серцебитті. Навіть візуально видно той самий нормальний звук серця, який має чіткий шаблон «луб даб, луб даб», це і є цикл нормального серцебиття.

На Рис. 2 теж видно те саме «бурчання», паузи між «луб» і «даб» нерівномірні, пікові значення та їх довжина різна.

На Рис. 3 теж приблизно проглядаються додаткові або пропущені серцеві скорочення, ті самі, «луб-луб-даб» або «луб-даб-даб». А на Рис. 4 явна демонстрація некоректних даних, які не відображають хоч якесь серцебиття.

На Рис. 5 більше пікових значень, тобто якраз і видно цей додатковий серцевий ритм.

5. Побудова моделей для класифікації і кластеризації часових рядів

Грунтуючись на багатьох дослідженнях було обрано по 2 [6], найкращих для цієї задачі, алгоритма кластеризації і класифікації, а саме методи K-means і DBSCAN для кластеризації часових рядів, а KNN і LSTM для класифікації.

Оскільки дані розмічені, а задачі кластеризації відносяться до так званого unsupervised learning (навчання без учителя), то відповідно видалено всі таргети для застосування кластеризації.

Для методу K-means основним параметром є кількість кластерів K. Для значення 5 результати були найкращі.

Для алгоритму DBSCAN аналогічно видалено таргети. Через не дуже великі об’єми даних DBSCAN навчається досить швидко, бо повільність навчання - є його відомим недоліком.

Алгоритм KNN ефективно відпрацював з гіперпараметрами за умовчанням, і як показав подальший аналіз для цих даних він є ефективним у співвідношенні якість/швидкості.

І в кінці проаналізовано найскладніший, але при цьому найефективніший підхід – рекурентну нейронну мережу LSTM.

Це багатошарова рекурентна мережа із функціями активації ReLU, а вже на вихідному шарі активація Softmax.

При використанні меншої кількості шарів модель мала меншу ефективність, але при цьому через велику загальну кількість параметрів LSTM навчалась дуже довго.

6. Аналіз отриманих результатів кластеризації часових рядів

Для усіх результатів ми виокремимо дві категорії: результати кластеризації і результати класифікації. Очевидно, що через різні підходи відповідно будуть різні метрики оцінювання якості алгоритмів і моделей.

Для алгоритмів DBSCAN і K-means обрано чотири метрики, які однаково можуть використовуватись як для K-середніх, так і для DBSCAN: silhouette_score, adjusted_rand_score, davies_bouldin_score та adjusted_mutual_info_score. Вони є широко використовуваними метриками для оцінки якості кластеризації в машинному навчанні.

- Silhouette Score - оцінює наскільки схожі між собою об'єкти в середині кластера та наскільки вони відрізняються від об'єктів в інших кластерах.
- Adjusted Rand Score - оцінює наскільки схожі кластеризації на істинні мітки. Значення 1 означає, що кластери повністю збігаються з істинними мітками.
- Davies-Bouldin Score - оцінює суміш внутрішньокластерної схожості та зовнішньокластерної відмінності для кожного кластера та робить узагальнення для всієї кластеризації. Чим нижче значення, тим краща кластеризація.
- Adjusted Mutual Information Score - оцінює наскільки взаємозв'язок між кластеризацією та істинними мітками відмінний від того, який очікується випадковим чином. Чим більше значення, тим краща кластеризація.

Одна з причин, чому метрики силуета, адаптована взаємна інформація, відстань Девіса-Болдуїна та адаптований рандомізований індекс чистоти підходять для кластеризації часових рядів, полягає в тому, що вони оцінюють якість кластерів, а не якість класифікації.

Ці метрики оцінюють, наскільки добре об'єкти в кожному кластері схожі між собою, і наскільки відмінні вони від об'єктів інших кластерів. Це дуже важливо для часових рядів, оскільки вони мають складну структуру та можуть мати різні форми. Метрики також дозволяють оцінити, чи є розділення на кластери зрозумілим та придатним для подальшого аналізу.

Таким чином, ці метрики допомагають зробити висновки про якість кластеризації та знайти найкращі параметри для алгоритмів кластеризації.

Проаналізуємо результати:

- Для K-середніх:
- Silhouette Score = 0.345
- Adjusted Rand Score = 0.006
- Davies-Bouldin Score = 0.876
- Adjusted Mutual Information Score = 0.049

Для DBSCAN:

- Silhouette Score = -0.6336115
- Adjusted Rand Score = 0.013803153550193674
- Davies-Bouldin Score = 1.339207713289357
- Adjusted Mutual Information Score = 0.037943281139379566

За результатами метрик можна сказати, що алгоритм K-means працює краще, ніж DBSCAN, для цього конкретного набору даних.

Він досягнув значень Silhouette Score близько до 0.34, що свідчить про те, що кластери добре відокремлені один від одного. Оцінка Adjusted Rand Score для K-means низька, а це може бути пов'язано з тим, що дані можуть містити шум, або кластеризація може бути досить складною.

Davies-Bouldin Score для K-means більше 0.87, що вказує на те, що кластери забезпечують добру відмінність один від одного, але можуть бути не такими оптимальними, як би ми хотіли. Adjusted Mutual Information Score для K-means також низький, що свідчить про низький рівень взаємозалежності між вхідними даними та отриманими кластерами.

З іншого боку, результати метрик для DBSCAN не такі високі [9]. Silhouette Score близький до -0.63, що показує, що кластери майже не відокремлені один від одного.

Оцінка Adjusted Rand Score для DBSCAN дещо вища, ніж для K-means, але все ще низька. Це означає, що кластери можуть мати різну кількість елементів і не відповідати оригінальним міткам. Davies-Bouldin Score для DBSCAN більше 1.33, що свідчить про те, що кластери не дуже відокремлені один від одного. Adjusted Mutual Information Score для DBSCAN також низький, що свідчить про низький рівень взаємозалежності між вхідними даними та отриманими кластерами.

Отже, можна зробити висновок, що алгоритм K-means більш ефективний для даного набору даних, оскільки досягнув більш високих значень метрик, але також варто зазначити, що результати метрик все рівно є високими, що свідчить про правильно обрані алгоритми та методи для кластеризації записів людського серцебиття.

7. Аналіз отриманих результатів класифікації часових рядів

Для оцінки алгоритму класифікації було обрано стандартні метрики для класифікації: accuracy, f1, precision, recall та матриця помилок (confusion matrix).

Метрики accuracy, f1, precision, recall та матриця помилок (confusion matrix) дуже добре підходять для задач класифікації часових рядів через те, що вони дають змогу оцінити якість класифікації на різних рівнях: загальну точність (accuracy), точність визначення позитивних класів (precision), точність визначення негативних класів (recall) та збіги та розбіжності в класифікації кожного з класів (confusion matrix). Враховуючи, що класифікація часових рядів є завданням з високою вимогою до точності та чутливості, використання цих метрик є дуже важливим для оцінки результатів роботи алгоритмів.

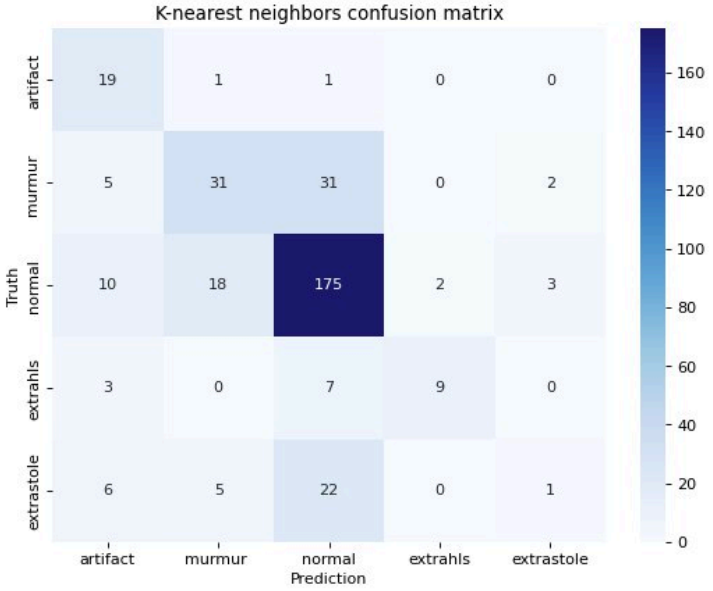


Рис. 6 – Матриця помилок для алгоритму KNN
Fig. 6 – Confusion Matrix for the KNN Algorithm

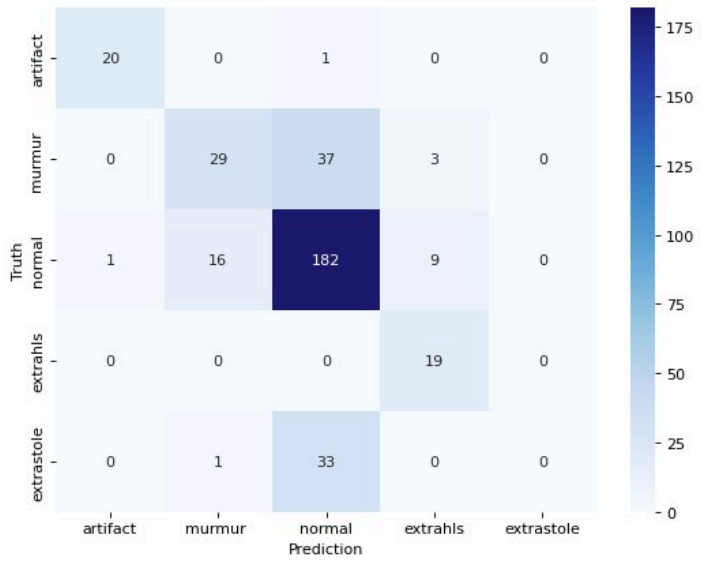


Рис. 7 – Матриця помилок для моделі LSTM
Fig. 7 – Confusion Matrix for the LSTM Model

Для KNN:

- Accuracy: 0.658,
- F1 score: 0.651,

- Recall: 0.658,
- Precision: 0.667

Для LSTM:

- Accuracy: 0.681,
- F1 score: 0.693,
- Recall: 0.681,
- Precision: 0.713

Дані результати метрик Accuracy, F1 score, Recall та Precision є високими для алгоритму K-Nearest Neighbors (KNN) у багатокласовій класифікації часових рядів.

Значення Accuracy вказує на те, що наш алгоритм правильно класифікував 65,8% зразків датасету, що є досить високим результатом. F1 score є гармонійним середнім між точністю (Precision) та повнотою (Recall), і також демонструє високу точність та повноту класифікації нашого алгоритму. Значення Precision та Recall становлять відповідно 0,667 та 0,658, що також є досить високими показниками.

Отже, високі результати метрик свідчать про ефективність використання алгоритму KNN для багатокласової класифікації часових рядів.

Як ми бачимо, результати метрик для LSTM не набагато краще, і може здатись, що алгоритм KNN, який простіший і швидший, краще підходить для даного набору даних, але це не так. Основна проблема полягає в тому, що комп’ютер, на якому проводились обчислення, не міг обробити велику кількість епох. Таким чином можна стверджувати, що отримані результати для LSTM можна значно покращити.

8. Висновки

1. Для кластеризації були використані алгоритми k-means та DBSCAN, що дозволило розділити записи на кілька категорій залежно від характеристик звуків. Для класифікації були використані алгоритми KNN та LSTM, що дозволило відрізнити звукові записи різних категорій та визначити, до якої категорії відноситься конкретний запис.

2. Отримані результати свідчать про ефективність використаних методів для аналізу звукових записів серцебиття людей та можуть бути використані для діагностики різних захворювань серця. Дослідження можуть бути продовжені з використанням інших алгоритмів та наборів даних з метою поліпшення точності класифікації та кластеризації.

3. Дослідження продемонструвало, що кластеризація та класифікація часових рядів з використанням алгоритмів k-means, DBSCAN, KNN і LSTM є ефективним методом для аналізу даних серцевих звуків.

4. Алгоритм k-means дозволяє кластеризувати дані серцевих звуків за їх характеристиками та дозволяє виявляти спільні риси між різними звуками. DBSCAN може бути корисним у виявленні аномальних звуків та відокремленні їх від нормальних. Алгоритм KNN забезпечує ефективну класифікацію звуків за їх характеристиками, тоді як LSTM може використовуватися для класифікації звуків на основі їх часових характеристик.

5. Отже, в даній роботі було успішно використано кластеризацію та класифікацію часових рядів для аналізу даних зі звуковими записами серцебиття людей, ділячи їх на категорії. Результати цієї роботи можуть бути корисними для медичних досліджень та можуть допомогти у розробці нових методів діагностики та лікування серцевих захворювань, а якщо їх продовжувати і розвивати, то і стати справді революційним методом у діагностуванні та виявленні захворювань серця і судинної системи.

Конфлікт інтересів

Автори повідомляють про відсутність конфлікту інтересів.

References

1. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
2. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10), 2222-2232. https://doi.org/10.1109/TNNLS.2016.2582924
3. Zhang, Z. (2004). Nearest neighbor search algorithms and applications. Springer. https://doi.org/10.1007/978-3-319-14717-8_39
4. Dasarathy, B. V. (1991). Nearest neighbor (NN) norms: NN pattern classification techniques. IEEE Computer Society Press.
5. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media. https://doi.org/10.1007/978-0-387-84858-7
6. Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), 645–678. https://doi.org/10.1109/TNN.2005.845141
7. Martin Ester, Jörg Sander (1996). "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications". *Data Mining and Knowledge Discovery*. 2 (2): 169–194. https://doi.org/10.1007/BF00457189
8. Hoang A.D., Bagnall A., Kaveh K., Chin-Chia M.Y., Zhu Y., Shaghayegh G., Chotirat A.R., Eamonn K. The UCR Time Series Archive URL: arxiv.org/abs/1810.07758
9. Schubert, E., Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (2017). "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN". *ACM Transactions on Database Systems (TODS)*, 42(3), 19. https://doi.org/10.1145/3068335
10. Kachanov Stanislav (2024) *Clustering and Classification of Time Series Data* (master diploma) V. N. Karazin Kharkiv National University

CLUSTERING AND CLASSIFICATION OF TIME SERIES SOUND DATA

Stanislav Kachanov[1], PhD Student; e-mail: staskachanov2000@gmail.com;
ORCID: https://orcid.org/0009-0002-6938-6717
Dmytro Vlasenko[1], senior lecturer of the Department of Theoretical and Applied Computer Sciences, PhD in mathematics; e-mail: vlasenkod@karazin.ua; ORCID: https://orcid.org/0009-0006-8780-2066
[1] *V. N. Karazin Kharkiv National University, Ukraine*

Manuscript was received March 17, 2024; Received after review April 19, 2024; Accepted May 20, 2024

Abstract. This scientific article addresses two critical tasks in data analysis—time series classification and clustering, particularly focusing on heart sound recordings. One of the main challenges in analyzing time series lies in the difficulty of comparing different series due to their variability in length, shape, and amplitude. Various algorithms were employed to tackle these tasks, including the Long Short-Term Memory (LSTM), KNN, recurrent neural network for classification and the K-means and DBSCAN methods for clustering. The study emphasizes the effectiveness of these methods in solving classification and clustering problems involving time series data containing heart sound recordings. The results indicate that LSTM is a powerful tool for time series classification due to its ability to retain contextual information over time. In contrast, KNN demonstrated high

accuracy and speed in classification, though its limitations became apparent with larger datasets. For clustering tasks, the K-means method proved to be more effective than DBSCAN, showing higher clustering quality based on metrics such as silhouette score, Rand score, and others. The data used in this research were obtained from the UCR Time Series Archive, which includes heart sound recordings from various categories: normal sounds, murmurs, additional heart sounds, artifacts, and extra systolic rhythms. The analysis of results demonstrated that the chosen classification and clustering methods could be effectively used for diagnosing heart diseases. Furthermore, this research opens up new opportunities for further improvement in data processing and analysis methods, particularly in developing new medical diagnostic tools. Thus, this work illustrates the effectiveness of machine learning algorithms for time series analysis and their significance in improving cardiovascular disease diagnosis.

Keywords: *time series classification, time series clustering, recurrent neural network, LSTM, KNN, K-means, DBSCAN, sound data analysis, heart sounds, machine learning, heart disease diagnosis*

Conflicts of Interest: the authors declare no conflict of interest.