

- [11] Y. Xu *et al.*, “Transformers in computational visual media: A survey,” *Comput. Vis. Media*, vol. 8, no. 1, pp. 33–62, Mar. 2022.
- [12] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, “Scaling vision transformers,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1204–1213.
- [13] X. Shi, H. Gu, and B. Yao, “Fuzzy Bayesian network fault diagnosis method based on fault tree for coal mine drainage system,” *IEEE Sensors J.*, vol. 24, no. 6, pp. 7537–7547, Mar. 2024.
- [14] P. D. M. Fernandez, F. A. G. Peña, T. I. Ren, and A. Cunha, “FERAtt: Facial expression recognition with attention net,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 837–846.
- [15] K. Han *et al.*, “A survey on vision transformer,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.
- [16] N. Aloysius and M. Geetha, “A review on deep convolutional neural networks,” in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Apr. 2017, pp. 0588–0592.
- [17] Y. Huang, F. Chen, S. Lv, and X. Wang, “Facial expression recognition: A survey,” *Symmetry*, vol. 11, no. 10, p. 1189, Sep. 2019.
- [18] A. S. Vyas, H. B. Prajapati, and V. K. Dabhi, “Survey on face expression recognition using CNN,” in *Proc. 5th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2019, pp. 102–106.
- [19] M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, and R. Segulier, “Learning vision transformer with squeeze and excitation for facial expression recognition,” *arXiv preprint*, arXiv:2107.03107, 2021.
- [20] F. Ma, B. Sun, and S. Li, “Facial expression recognition with visual transformers and attentional selective fusion,” *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1236–1248, Oct. 2021.
- [21] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, “ViTFER: Facial emotion recognition with vision transformers,” *Appl. Syst. Innov.*, vol. 5, no. 4, p. 80, Aug. 2022.

УДК 004.932.2:004.932.72:004.91

ПОРІВНЯЛЬНА ХАРАКТЕРИСТИКА НЕЙРОННИХ АРХІТЕКТУР CNN ТА RNN/LSTM ДЛЯ ЗАДАЧІ КЛАСИФІКАЦІЇ МУЗИЧНИХ ЖАНРІВ

Ярченко Д. С., Пархоменко О. Ю.

(arsenkokod@gmail.com, parkhomenko@mnaeu.edu.ua)

Миколаївський національний аграрний університет (Україна)

Проведено порівняльний аналіз двох основних типів нейронних мереж для задачі класифікації музичних жанрів: згорткових нейронних мереж (CNN) та рекурентних нейронних мереж (RNN/LSTM). Розглянуто принципи функціонування обох архітектур та їх застосування до обробки аудіосигналів. Показано, що CNN ефективно виділяють локальні спектральні ознаки з коротких аудіофрагментів, забезпечуючи високу швидкість обробки та точність понад 85% на стандартних датасетах. LSTM-мережі демонструють перевагу при аналізі довгих композицій завдяки здатності моделювати часову динаміку та враховувати довгострокові залежності між музичними фразами. Обґрунтовано доцільність використання гібридних архітектур, що поєднують CNN для екстракції ознак зі спектрограм та LSTM для моделювання їх часової еволюції, що дозволяє досягти найкращих результатів класифікації музичних жанрів.

Класифікація музичних жанрів є важливим завданням у сфері обробки аудіосигналів, оскільки дозволяє автоматизувати процеси організації, індексації та рекомендації музичного контенту в стрімінгових платформах і мультимедійних системах. Із розвитком глибинного навчання з'явилися нові методи на основі нейронних мереж, які значно підвищили точність розпізнавання жанрів. Однак існує проблема вибору оптимальної архітектури нейронної мережі залежно від характеристик аудіосигналу та специфіки застосування. Два домінуючі підходи – згорткові нейронні мережі (CNN) та рекурентні нейронні мережі (RNN/LSTM) – базуються на різних принципах і мають відмінні переваги та обмеження, що потребує їх детального порівняльного аналізу.

Метою роботи є порівняльний аналіз архітектур CNN та RNN/LSTM для задачі класифікації музичних жанрів та визначення найбільш оптимального підходу. Для досягнення мети вирішувалися наступні завдання: 1) дослідити принципи функціонування CNN та їх застосування до обробки спектрограм аудіосигналів; 2) проаналізувати можливості LSTM-мереж у моделюванні часової динаміки музичних композицій; 3) порівняти ефективність обох архітектур; 4) обґрунтувати доцільність використання гібридних підходів для досягнення максимальної точності класифікації.

Архітектура CNN. Згорткові нейронні мережі були розроблені першочергово для аналізу зображень, однак завдяки універсальності архітектури вони виявилися надзвичайно ефективними для багатьох типів вхідних даних, що мають просторову структуру, зокрема аудіо у формі спектрограм. Аудіосигнал, який є одночасно і часовою, і частотною інформацією, перетворюється в спектрограму – графічне представлення енергетичного розподілу сигналу в часі й частоті. Згорткові шари в CNN здатні автоматично знаходити важливі локальні патерни, наприклад, характерні переходи між нотами, ритмічні удари, специфічні особливості тембру інструментів тощо.

Ці патерни особливо важливі для розпізнавання музичних жанрів, оскільки більшість жанрів мають унікальні темброві та ритмічні особливості, які проявляються у коротких відрізках сигналу. Завдяки своїй архітектурі, CNN можуть ефективно виділяти ознаки на різних рівнях абстракції – від простих локальних фільтрів до комплексних концептів, які узагальнюють інформацію про жанр. Це дозволяє суттєво зменшити потребу в ручному формуванні ознак, характерному для класичних підходів [1].

Крім того, CNN демонструють високу продуктивність і масштабованість, оскільки більшість операцій можна реалізовувати паралельно. Це робить такі мережі зручними для обробки великих аудіодатасетів, де потрібна як швидка обробка, так і висока точність. Водночас, через обмежену здатність до моделювання послідовності, CNN мають певні труднощі з аналізом тривалої часової структури музичного твору, особливо якщо жанрові особливості проявляються не в коротких фрагментах, а у повній композиційній побудові.

Архітектура RNN/LSTM. Рекурентні нейронні мережі (RNN) призначені для роботи з послідовними даними, де порядок елементів є важливим. Вони мають внутрішню пам'ять, яка дозволяє враховувати попередній контекст при обробці поточного елементу послідовності. Проте класичні RNN мають суттєві обмеження, зокрема проблему затухання або вибуху градієнта, що ускладнює їх застосування при довгих послідовностях. Саме тому значно більш ефективною виявилася модифікація – LSTM (Long Short-Term Memory) [2]. (рисунк.1)

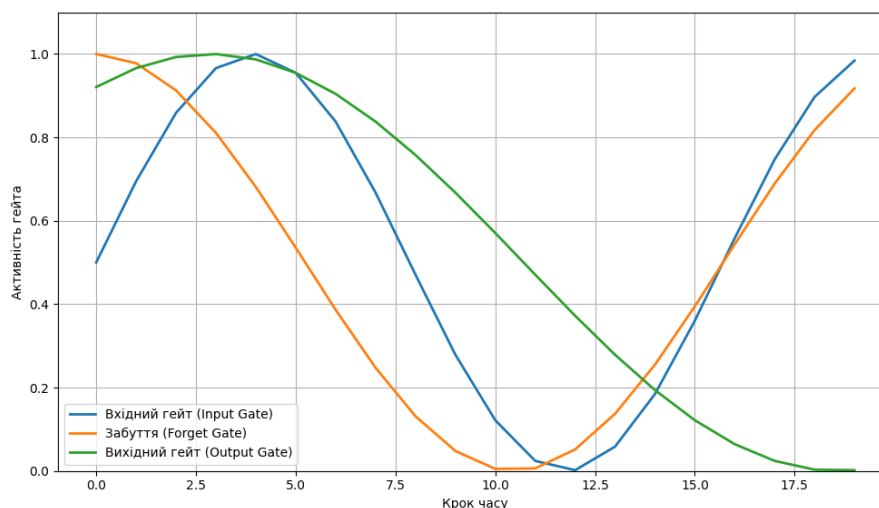


Рис. 1. Ілюстрація роботи гейтів LSTM у часі

LSTM має спеціальні механізми – так звані "вентилі" (гейти), які дозволяють контролювати потік інформації у внутрішній пам'яті: що запам'ятати, що забути, а що передати далі. Завдяки цьому модель може ефективно "пам'ятати" інформацію на великій кількості кроків, що є надзвичайно важливим при аналізі музичних творів, де послідовність розвитку музики має визначальне значення для її жанрової ідентифікації.

Наприклад, у класичній музиці чи джазі структура композиції часто є важливою ознакою жанру – з характерними модуляціями, тематичним розвитком, імпровізаційними вставками або специфічною формою. У таких випадках CNN можуть не зафіксувати глобальні залежності, тоді як LSTM здатна вловити зв'язки між музичними фразами, що знаходяться на значній відстані одна від одної у часовому вимірі. Це дає змогу моделі робити обґрунтовані висновки щодо жанрової належності композиції.

Однак застосування LSTM супроводжується певними обмеженнями. Зокрема, навчання таких мереж відбувається повільніше, ніж CNN, через необхідність послідовної обробки вхідних даних. Також, через складнішу структуру, вони є більш ресурсоемними і вимагають ретельної оптимізації. Ці чинники можуть стати перешкодою для використання LSTM у реальному часі або на пристроях з обмеженими обчислювальними ресурсами [3].

Порівняльні експерименти та гібридні підходи. У багатьох експериментальних дослідженнях, які проводилися на публічних датасетах (наприклад, GTZAN, FMA, MagnaTagATune, UrbanSound8K), CNN показують високі результати в задачах класифікації музичних жанрів, особливо при роботі з короткими аудіофрагментами (наприклад, 10–30 секунд). В таких умовах перевага локального аналізу та швидкість обробки виявляються вирішальними. У середньому, згорткові мережі досягають точності понад 85% без складної попередньої обробки сигналу.

З іншого боку, LSTM демонструють вищу точність у випадках, коли використовуються цілі музичні треки або великі фрагменти з послідовною подачею вхідних ознак. У таких сценаріях мережа здатна накопичувати контекст і використовувати довготривалу пам'ять для обґрунтованого прийняття рішень. (рисунок.2)

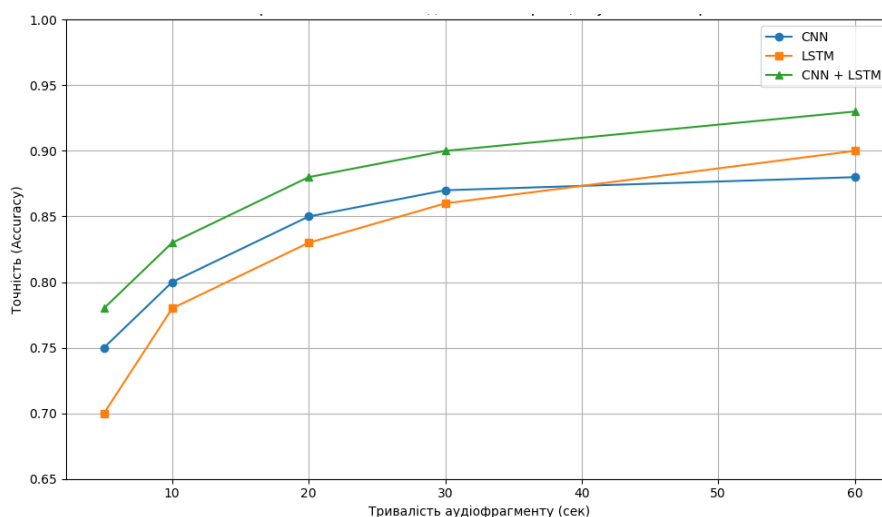


Рис. 2. Порівняння точності моделей класифікації музичних жанрів

Найбільш ефективним підходом вважається **комбінування CNN та LSTM** в одній архітектурі. Такий гібрид дозволяє спершу екстрагувати спектральні ознаки за допомогою згорткових шарів, а потім передати ці ознаки в LSTM-шар для моделювання їх динаміки у часі. У результаті, модель отримує доступ як до локальних (тембрових, ритмічних), так і до глобальних (структурних) характеристик композиції. Цей підхід довів свою ефективність у численних дослідженнях і продовжує залишатися одним із найпопулярніших у сучасних системах музичної класифікації.

У результаті проведеного дослідження вирішено поставлені завдання та отримано наступні результати. По-перше, встановлено, що CNN ефективно виділяють локальні спектральні ознаки з коротких аудіофрагментів (10-30 секунд), забезпечуючи високу швидкість обробки та точність понад 85% на стандартних датасетах завдяки здатності автоматично знаходити характерні темброві та ритмічні патерни. По-друге, доведено, що LSTM-мережі демонструють вищу ефективність при аналізі довгих композицій завдяки механізмам внутрішньої пам'яті, що дозволяє моделювати часову динаміку та враховувати довгострокові залежності між музичними фразами, які є критичними для жанрової ідентифікації складних композицій. По-третє, на основі аналізу експериментальних досліджень на публічних датасетах обґрунтовано, що оптимальний результат

класифікації музичних жанрів досягається за умови поєднання обох підходів у гібридній архітектурі, де CNN екстрагують спектральні ознаки, а LSTM моделює їх часову еволюцію, що дозволяє максимально повно охопити як локальні, так і глобальні характеристики музичного сигналу.

Список використаної літератури

- [1] V. Kazymyr, A. Rohovenko та O. Karas, “Application of Artificial Neural Networks for Audio Classification on an Embedded Platform”, Tech. sci. technol., № 4 (38), с. 134–149, груд. 2024. Дата звернення: 22 жовт. 2025. [Онлайн]. Доступно: [https://doi.org/10.25140/2411-5363-2024-4\(38\)-134-149](https://doi.org/10.25140/2411-5363-2024-4(38)-134-149)
- [2] Д. Є. Пугач, «Розробка голосової системи активації з використанням нейронних мереж», Кваліфікаційна робота магістра, спеціальність 121 Інженерія програмного забезпечення, наук. керівник С. М. Гребенюк, Запоріжжя: ЗНУ, 2024, 54 с. Дата звернення: 22 жовт. 2025. [Онлайн]. Доступно: <https://dspace.znu.edu.ua/jspui/bitstream/12345/24458/1/Пугач.pdf>
- [3] М. С. Штойко, «Метод пояснення результатів задач класифікації за моделями глибокого навчання засобами машинного навчання», Кваліфікаційна робота магістра, спеціальність 122 Комп’ютерні науки, Хмельницький: Хмельницький нац. ун-т, 2024, 108 с. Дата звернення: 22 жовт. 2025. [Онлайн]. Доступно: <https://elar.khmnu.edu.ua/items/82fce3e3-c333-4f4e-97a7-4687f90f8558>