УДК 78.02:[004.81:159.953.5
DOI https://doi.org/10.24919/2308-4863/86-3-11

Юрій СМАЛІУС,
orcid.org/0009-0004-1264-9226
аспірант
Національної академії керівних кадрів культури і мистецтв
(Київ, Україна) dmz2123.ysmalius@dakkkim.edu.ua

НЕЙРОННІ МЕРЕЖІ ДЛЯ СТВОРЕННЯ МУЗИЧНИХ КОМПОЗИЦІЙ: ПРИНЦИПИ РОБОТИ ТА ПРАКТИЧНЕ ЗАСТОСУВАННЯ

У статті охарактеризовано глибокі нейронні мережі – DeepBach, MuseGAN, WaveNet (DeepMind) Jukebox (OpenAI), Noise2Music, Riffusion та інших, що дозволяють генерувати музику у різних жанрах із реалістичним звучанням. Методологія роботи включає аналітичний, структурно-функціональний, узагальнюючий методи дослідження. Наукова новизна роботи. На основі вивчення нейронних мереж – DeepBach, MuseGAN, WaveNet (DeepMind) Jukebox (OpenAI), Noise2Music, Riffusion було зроблено оцінку потенціалу застосування таких сучасних інструментів аудіовиробництва для створення музичних композицій, як таких, що можуть бути відтворені за допомогою додаткового обладнання, маючи, відповідно, більший потенціал для редакції та вдосконалення, так і таких, що є довершеною цифровою репрезентацією аудіо сигналу музичної композиції. Висновки. Досліджено основні типи нейронних мереж глибокого навчання в контексті автономної генерації музичних композицій за текстовим запитом. Дослідження розділене на два основних напрямки: символьна генерація (музичний текст для відтворення) та моделі, що генерують цифрові репрезентації аудіосигналівю. Моделі символьної генерації демонструють значний прогрес у завданнях гармонізації, побудови логічних музичних послідовностей і стилізації під різні жанри. Розглянуто сильні та слабкі сторони застосування символьної генерайії у аудіовиробництві. В свою чергу аналіз побидови та характеру роботи моделей, які безпосередньо генерують цифрові репрезентації аудіосигналів, демонструють безпрецедентні успіхи дифузійних моделей, які дозволяють отримувати реалістичніші звучання та враховувати семантичні аспекти тексту й культурного контексту. Окрім цього, окреслено певні технологічні недоліки вихідного матеріалу, що може включати (але не обмежується) артефакти звучання, спотворення, надмірна складність та ресурсоємність процесів навчання моделей, що ускладнюють їх широке використання. Надано оцінку потенціалу застосування таких технологій у музичному та аудіовиробництві.

Ключові слова: *штучний інтелект, нейронні мережі глибокого навчання, генерована музика, LSTM, дифузні моделі, алгоритмічні моделі, символьна генерація, генерація цифрового представлення аудіо сигналів.*

Yurii SMALIUS,
orcid.org/0009-0004-1264-9226
PhD Candidate
National Academy of Culture and Arts Management
(Kyiv, Ukraine) dmz2123.ysmalius@dakkkim.edu.ua

NEURAL NETWORKS FOR CREATING MUSICAL COMPOSITIONS: PRINCIPLES OF OPERATION AND PRACTICAL APPLICATION

The article describes deep neural networks – DeepBach, MuseGAN, WaveNet (DeepMind) Jukebox (OpenAI), Noise2Music, Riffusion and others, which allow generating music in various genres with realistic sound. The methodology of the research incorporates analytical, structural-functional, and generalization approaches. The study's methodology encompasses analytical, structural-functional, and generalizing methods. Scientific novelty. On the basis of examination of neural networks such as DeepBach, MuseGAN, WaveNet (DeepMind), Jukebox (OpenAI), Noise2Music, and Riffusion, an assessment was conducted regarding the potential application of these contemporary audio production tools for creating musical compositions. This includes compositions intended for reproduction using additional equipment, offering greater potential for editing and refinement, as well as compositions that represent digital audio signals. Conclusions. The study investigated major types of deep learning neural networks in the context of autonomous music generation based on textual queries. The research is divided into two principal directions: symbolic generation (musical notation for playback) and models generating digital representations of audio signals. Symbolic generation models demonstrate considerable progress in tasks such as harmonization, logical musical sequence construction, and stylization into various genres. The strengths and weaknesses of symbolic generation in audio production were examined. In turn, the analysis of the architecture and functioning of models directly generating digital audio representations revealed the unprecedented success of diffusion models, enabling more realistic sound generation while also considering semantic elements of the text and cultural contexts. Additionally, certain technological shortcomings of the output material were highlighted, including but not limited to auditory artifacts, distortions, excessive complexity, and the resource-intensive nature of model training processes, complicating their widespread adoption. The article provides an evaluation of the potential applications of these technologies in music and audio production.

Key words: *artificial intelligence, deep learning neural networks, generated music, LSTM, diffusion models, algorithmic models, symbolic generation, digital audio signal generation.*

Постановка проблеми. Сучасний розвиток штучного інтелекту (ШІ) значно вплинув на сферу музичної творчості, відкриваючи нові можливості для композиторів, виконавців і дослідників. Стрімкий розвиток технологій глибокого навчання, особливо після 2010-х років, зробив можливим використання потужних нейронних мереж у процесі виробництва музики. ШІ у музиці працює на різних ключових рівнях. Можемо виокремити роботу з нотними партитурами, послідовностями акордів, мелодичними та ритмічними паттернами. Нейронні мережі можуть аналізувати великі музичні архіви та генерувати нові композиції у стилі конкретних композиторів або жанрів. Але вивчення таких моделей, зокрема DeepBach, MuseGAN, WaveNet (DeepMind) Jukebox (OpenAI), Noise2Music, Riffusion тощо залишається поза увагою дослідників і стає актуальною проблематикою для музикознавців.

Аналіз досліджень. У статті «Застосування та прогрес штучного інтелекту в музичній генерації: огляд» (Chen, Huang і Gou, 2024) зроблено акцент на узагальненні новітніх досягнень у музичній генерації. Дослідники поділяють алгоритми на три основні категорії (символьні, аудіо й гібридні) та детально аналізують типові архітектури кожної. Також наводиться аналітичний зріз практичних застосувань – від музикознавчих досліджень до інтерактивних творчих платформ, орієнтованих на реалії сучасних музичних індустрій. Дослідження «Технології штучного інтелекту для створення поп-музики: погляд із позиції реального музичного продакшену» (Zhang, Yan і Briot, 2023) має прикладну спрямованість і націлене на аспекти створення поп-музики. Автори описують етапи типового «виробничого ланцюга» у поп-композиції: від генерації мелодії та тексту до аранжування та остаточного аудіорендерінгу. Важливим є те, що приділяється увага саме «промисловому» підходу, коли інтеграція ШІ-алгоритмів відбувається в реальних студійних процесах. Висновки вказують, що AI-технології дають змогу прототипувати пісні швидше й отримувати варіативні стилі, але потребують удосконалення для досягнення рівня професійної виразності та деталізації звучання. Робота «DeepBach: керована модель для генерації хоралів Баха» (Hadjeres, Pachet і Nielsen, 2017) присвячена моделі DeepBach, яка добре відома в контексті автоматизованої гармонізації хоралів Й. С. Баха. Ключова ідея полягає в застосуванні псевдо-ґіббсівського семплування для поступового «виправлення» нот у кількох голосах, що контрастує з типовою послідовною генерацією «зліва направо». Така модель дає змогу користувачеві керувати процесом і досягати результатів, які часто важко відрізнити від оригінальної бахівської стилістики. Цей підхід довів ефективність нетривіальних архітектур і важливість інтерактивної взаємодії композитора з моделлю. У статті «MusicLM: генерація музики з тексту» (Agostinelli et al., 2023) продемонстровано масштабне зрушення в генерації безпосередньо цифрового представлення аудіосигналу. Модель ґрунтується на каскадному перетворенні (audio tokens), що покращує довготривалу узгодженість і дає змогу формувати музику, враховуючи стислий текстовий опис. Авторами зазначено високий рівень стилістичної відповідності та наявність якісних тембральних переходів, хоча наголошується й на існуючих викликах. Важливим дослідженням, що розкриває принципи побудови та функціонування моделей ШІ, які генерують цифрове представлення аудіосигналів, також є стаття «Noise2Music: текстово-зумовлена генерація музики з використанням дифузійних моделей» (Huang, Park та ін., 2023), що описує дифузійнійну систему Noise2Music, показує дієвість каскадних моделей у синтезі 30-секундних уривків на основі тексту. Підхід засновано на проміжному поданні вихідних даних низької якості (спектрограма чи аудіо), яке потім проходить послідовне «уточнення» аж до цілком реалістичного звучання. Дослідники також наголошують на залученні великих мовних моделей для попереднього генерування анотацій, що дає можливість масштабованого навчання.

Мета статті полягає у дослідженні нейронних мереж глибокого навчання – DeepBach, MuseGAN, WaveNet (DeepMind) Jukebox (OpenAI), Noise2Music, Riffusion та інших, що дозволяють генерувати музику у різних жанрах із реалістичним звучанням.

Виклад основного матеріалу. Вивчення основної проблематики стосується двох напрямів: символьної генерації та генерації відтворюваних репрезентацій цифрових аудіосигналів. Перший із них – символьна генерація музики, що базується на застосуванні ШІ для формування дискретних музичних представлень, таких як MIDI-файли, нотні записи чи програмовані патерни. Основна ідея цього підходу полягає у опануванні моделлю музичних структур, прогресій гармонії, мелодійних ліній та ритмічних патернів задля формування композицій із чіткою логікою та структурною організацією форми, які відповідають встановленим традиціям і стилям написання музики (Edwards. 2011: 58). Оскільки такі моделі працюють з дискретними даними (цифровими нотами), результати генерації самі по собі не є цифровим представлен-

ням аудіо хвиль і потребують подальшого «озвучення» за допомогою синтезаторів, віртуальних інструментів, семплерів, драм-машин тощо.

Однією з найвідоміших ранніх моделей глибокого навчання для символічної музики є DeepBach (Hadjeres et al., 2017). Її предметна область доволі вузька, але показова – генерація чотириголосних гармонізацій хоралів в стилі Й. С. Баха. DeepBach позиціонується як стилеспецифічна модель: навчившись на базі 389 хоралів Баха, вона здатна продукувати нові хоральні гармонізації, які за стилем імітують оригінал (Hadjeres G., Pachet F., Nielsen F., 2017).

Важливою перевагою DeepBach є можливість інтерактивного керування генерацією. Модель задумана як «керована» (steerable): користувач може накладати певні обмеження на генерований твір – наприклад, зафіксувати кілька нот, ритмічний малюнок або гармонійну послідовність у каденції (Hadjeres G., Pachet F., Nielsen F., 2017).

Якість згенерованих DeepBach хоралів було оцінено як об'єктивно, так і за допомогою слухацького тесту. Об'єктивно модель практично не робить грубих помилок гармонії: завдяки навчанню на стилізованих та обраних даних вона дотримується правил класичного чотириголосся (уникнення паралельних квінт тощо) майже бездоганно. Суб'єктивну переконливість перевіряли у форматі тесту «Бах чи комп'ютер», де слухачам пропонували короткі фрагменти гармонізації і вимагали вгадати, чи написані вони людиною (Бахом) або ж згенеровані моделлю. У тестуванні взяли участь 1272 респонденти трьох рівнів музичної підготовки (нефахівці, любителі та професіонали). Результати показали, що DeepBach може переконати слухачів приблизно в половині випадків: близько 50% прослуховувань створених генерацій були оцінені як справжні бахівські твори (Hadjeres G., Pachet F., Nielsen F., 2017).

Принцип роботи цієї та подібних моделей базується на основі LSTM, що розшифровується як Long Short-Term Memory (довготривала короткострокова пам'ять) – це тип рекурентних нейронних мереж (RNN), розроблений для подолання проблеми збереження довгострокових залежностей у послідовних даних. Принцип функціонування LSTM полягає у використанні спеціалізованих елементів, що називаються «комірками пам'яті», які здатні зберігати інформацію протягом тривалого періоду. Кожна така комірка має кілька «гейтів» (воріт): забуваючий гейт (Forget Gate), що визначає, яку частину інформації попереднього стану слід «забути»; вхідний гейт (Input Gate), що регулює, яку нову інформацію включити до стану комірки; вихідний гейт (Output Gate), що контролює, яку інформацію передати як вихід моделі. Цікаво, втім, що саме модель DeepBach архітектурно не просто є черговою послідовною LSTM-моделлю – автори називають її графічною моделлю із застосуванням алгоритму псевдо-Гіббсівського вибіркового семплування (Hadjeres G., Pachet F., Nielsen F., 2017).

На відміну від традиційного підходу з генерацією нота за нотою зліва направо, DeepBach дозволяє заповнювати музичний фрагмент поступово, перегенеровуючи ноти в різних голосах до отримання узгодженого результату. Фактично модель складається з ансамблю двонапрямлених LSTM, що вчаться передбачати ноти кожного голосу на основі оточення, і генерує весь чотириголосний твір шляхом ітеративного оновлення нот на основі ймовірностей (Hadjeres G., Pachet F., Nielsen F., 2017).

Такий підхід добре підходить для задач із гармонізації, де важливо враховувати контекст як ліворуч, так і праворуч від поточної ноти (наприклад, басову лінію як опору і мелодію хоралу задають наперед, а середні голоси заповнюють).

Більші можливості у виявленні довготривалих залежностей та послідовностей демонструють моделі генерації символьної музики на базі трансформерів. Це нейронні мережі, які використовують архітектуру трансформерів для побудови музичних послідовностей, представлених у вигляді символьних даних. Основний принцип цих моделей ґрунтується на механізмі самоуваги (self-attention), що дозволяє одночасно аналізувати всі елементи вхідної послідовності та встановлювати між ними взаємозв'язки незалежно від їх розташування. Завдяки цьому трансформер здатний ефективно відображати як короткострокові, так і довгострокові залежності в даних, що є важливим для збереження структурної цілісності музичних творів (Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., & Yang, Y.-H., 2017). До того ж, застосування позиційних кодувань забезпечує збереження порядку елементів, що є критичним для правильного інтерпретування ритмічних і мелодійних структур. Такі моделі дозволяють генерувати композиції з чіткою логічною структурою, відтворюючи різноманітні музичні стилі та жанри.

Як приклад можна навести модель Pop Music Transformer (Huang та Yang, 2020), яка інтегрує механізми самоуваги та архітектуру трансформерів для досягнення значних покращень у генерації послідовностей в контексті поп-музики. В експериментах Pop Music Transformer порівнювався з базовою моделлю без спеціального подання. Було показано, що з уведенням метричної інфор-

мації трансформер генерує поп-музику (цільовий жанр – сольні фортепіанні композиції у стилі поп) з набагато кращою ритмічною структурою (Huang, Y.-S., & Yang, Y.-H., 2020: 1180). Зокрема, зменшилась частка «випадкових» зсувів і пауз, музика краще узгоджувалася з розмірами і не втрачала битого пульсу. При цьому збереглася і загальна сила трансформера – моделювання довготривалих залежностей: модель може генерувати до хвилини музики, що має осмислений розвиток (Huang, Y.-S., & Yang, Y.-H., 2020: 1182–1183). Таким чином, імплементація музикознавчого знання у форму вхідних даних показала себе дієвим способом підвищити ефективність моделі.

Також варта уваги модель MuseGAN. Це система генерації багатодоріжкової символьної музики на основі генеративних змагальних мереж (GAN). Принцип роботи таких моделей ґрунтується на змагальному навчанні двох мереж: генератора (G) та дискримінатора (D). Генератор синтезує символьні музичні послідовності, які мають імітувати статистичні властивості реальних даних, тоді як дискримінатор відрізняє генеровані послідовності від реальних зразків. Обидві мережі оптимізуються за схемою мінімакс, де генератор мінімізує функцію втрат, а дискримінатор – максимізує її. Цей процес забезпечує поступове підвищення якості синтезованих творів при кожній новій ітерації зі збереженням та структури гармонії та цілісності форми багато доріжкових композицій. Це надає потенціал для створення більш складних та досконалих музичних текстів.

Хоча MuseGAN генерує лише короткі уривки (чотиритактові патерни), її важливий внесок – показати принципову можливість створення поліфонічної багатошарової музики засобами глибинного навчання. Модель довела, що GAN може оперувати складними музичними структурами (кілька паралельних послідовностей нот) і видавати результат, який слухачі оцінюють цілком позитивно. Згенеровані фрагменти, за відгуками, звучать як осмислені музичні ідеї – у них є відчуття гармонії, ритму, початок і завершення фраз (Edwards, M., 2011: 58–59). Водночас, автори самокритично зауважують, що в художньому відношенні MuseGAN ще далекий від людського рівня: навіть найкращі його зразки все ж менш цікаві й різноманітні, ніж те, що може придумати професійний композитор (Edwards, M., 2011: 59–60). Проте на час оприявлення (2018) модель MuseGAN стала значним кроком вперед, адже продемонструвала роботу генеративної мережі на багатоголосому фактурному матеріалі поп-музики.

Генерація музики у аудіо форматі безпосередньо спрямована на створення дискретної репрезентації аудіосигналів, що можуть бути як негайно відтворені апаратними чи програмними засобами, так і використані для подальшої роботи в музичному продакшені. Завдяки цьому підходу, що є наближеним до етапів запису та мікшування у музичному виробництві, можливо отримувати музичний контент доволі високого рівня реалістичності. Варто зазначити, що описані вище типи моделей, які використовуються для символьної генерації, великою мірою описують і механізми роботи аудіо генеруючих моделей. Але окремо слід позначити такий тип моделей, як дифузійні.

У дослідженні «MusicLM: Generating Music From Text» (Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., & Frank, 2023) продемонстрували, що дифузійні моделі особливо вдало враховують культурний контекст у музичних творах. При порівнянні генерації музики з однаковим жанром, але різними мовними параметрами, модель продукувала суттєво відмінні композиції. Наприклад, при запиті «традиційна пісня, мова китайська» модель генерувала твори, де були присутні типові для української музики інтервали та гармонічні послідовності, у той час як за запитом «традиційна пісня, мова іспанська», результати містили характерні фламенко-подібні елементи (Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., & Frank, 2023).

У музиці дифузійні методи теж почали давати результати. Аудіосигнал можна уявити як послідовність, і модель навчається крок за кроком покращувати його якість: з чистого шуму – до впізнаваних музичних структур. Складність у тому, що музика має високу частоту дискретизації (десятки тисяч точок на секунду), тому генерація потребує багато ітерацій. Сучасні рішення застосовують оптимізації, наприклад генерують не сам сигнал відразу, а проміжне представлення.

У роботі «Noise2Music: Text-conditioned Music Generation with Diffusion Models» було запропоновано каскад дифузійних моделей для генерації 30-секундних музичних кліпів з текстового опису (Huang, Q., Park, D. S., Wang, T., Denk, T. I., Ly, A., Chen, N., Zhang, Z., Zhang, Z., Yu, J., Frank, C., Engel, J., Le, Q. V., Chan, W., Chen, Z., & Han, W, 2023). Перша модель (генератор) створює проміжне низькоякісне представлення музики (спектрограму невеликого розміру або аудіосигнал низької частоти дискретизації на рівні 3 кГц), яке відповідає тексту, а друга модель (каскадер) перетворює це представлення на більш прийнятний за якістю (аудіо 24 кГц, 30 секунд тривалістю) аудіосигнал (Huang, Q., Park, D. S., Wang, T., Denk, T. I.,

Ly, A., Chen, N., Zhang, Z., Zhang, Z., Yu, J., Frank, C., Engel, J., Le, Q. V., Chan, W., Chen, Z., & Han, W, 2023). Цікаво, що автори випробували дві форми проміжного представлення – спектрограму та низькоякісне аудіо – і обидва підходи дали змогу отримати музичні уривки, які чітко відображають ключові елементи текстового запиту: жанр, темп, інструменти, настрій і навіть епоху (Huang, Q., Park, D. S., Wang, T., Denk, T. I., Ly, A., Chen, N., Zhang, Z., Zhang, Z., Yu, J., Frank, C., Engel, J., Le, Q. V., Chan, W., Chen, Z., & Han, W, 2023). Це свідчить, що дифузійна модель може «розуміти» та втілювати у звуці доволі тонкі семантичні нюанси опису. Дифузійні моделі для музики часто поєднуються з іншими техніками. Наприклад, дослідники з Google використали попередньо навчені великі мовні моделі для генерації текстових анотацій до аудіоданих, щоб отримати великі пари «опис–музика» для тренування дифузійної моделі (Huang, Q., Park, D. S., Wang, T., Denk, T. I., Ly, A., Chen, N., Zhang, Z., Zhang, Z., Yu, J., Frank, C., Engel, J., Le, Q. V., Chan, W., Chen, Z., & Han, W, 2023).

WaveNet – нейромережа, що стала однією із перших вдалих спроб втілення інноваційних підходів у галузі нейромережевого генерування аудіо. Ця модель заслуговує згадки, оскільки багато пізніших робіт відштовхувалися від ідей, що були її частиною. У моделі WaveNet використовується глибинна згорткова нейромережа з дилатацією для моделювання ймовірнісного розподілу наступного семплу звуку на основі попередніх семплів (Chen, Y., Huang, L., & Gou, T., 2024).

Модель продемонструвала напрочуд реалістичні результати у синтезі мовлення, відтворюючи природні інтонації і тембр людського голосу, а також фортепіанних звуків. Для музики WaveNet також може застосовуватися, але через свою авторегресивну природу генерує звук посемплово, що для довгих музичних відрізків надто повільно і громіздко. Тим не менш, у 2020 році з'явився проєкт **Jukebox** від OpenAI. Це модель, яка частково базувалася на ідеях WaveNet, але вводила кількарівневе стиснення даних через VQ-VAE, щоб зменшити послідовність для генерування (Civit, M., Civit-Masot, J., Cuadrado, F., & Escalona, M. J., 2022). Модель Jukebox спромоглася генерувати хвилини музики у вибраному стилі і навіть з імітацією голосу та тексту пісні (Civit, M., Civit-Masot, J., Cuadrado, F., & Escalona, M. J., 2022). Його принцип роботи: спершу закодувати аудіо у дискретні кодеки (семпли не напряму, а через «пакети» інформації), а потім генерувати ці кодеки трансформером, подібно до того як генерують символи тексту. В результаті Jukebox міг створювати повноцінні аудіотреки з різними інструментами і співом. Важливо зазначити, що спів у Jukebox – це радше імітація вокального тембру: модель не володіла справжнім мовним розумінням, тому тексти пісень виходили беззмістовним набором фонетичних відтворень. Та все ж, стилістично Jukebox міг зорієнтуватися (напр., «пісня в стилі Елвіса Преслі») і видати відповідний результат (Civit, M., Civit-Masot, J., Cuadrado, F., & Escalona, M. J., 2022).

Одним із обмежень прямої аудіогенерації, яке проявилося в Jukebox і подібних підходах, є артефакти у звуці. Навіть якщо на коротких відрізках звук здається реалістичним, при довгому прослуховуванні з'являються помітні повтори, цифровий шум, спотворення чи інші нехарактерні для живого запису деталі. Дослідники відзначають, що аудіоряди, згенеровані на основі нестиснутих WAV-файлів (наприклад, Jukebox чи WaveNet), наразі містять помітні артефакти, які видають їх штучне походження (Civit, M., Civit-Masot, J., Cuadrado, F., & Escalona, M. J., 2022). Людський слух дуже чутливий до таких дрібниць, тому подібні моделі поки що складно використати безпосередньо у професійному музичному виробництві (Civit, M., Civit-Masot, J., Cuadrado, F., & Escalona, M. J., 2022). Натомість, символьні моделі (MIDI) у цьому аспекті мали перевагу: їхній «вивід» міг бути відредагований та відтворений на якісних семплах людиною, що гарантувало чистий звук. Зважаючи на це, при використанні ШІ у аудіовиробництві надавалась перевага символічній генерації з подальшим рендерінгом у аудіо за допомогою наявних бібліотек звуків.

Окремо варто відзначити креативний підхід Riffusion (2022), який продемонстрував можливість генерувати музику за допомогою вже наявних моделей дифузії зображень. Riffusion навчали на зображеннях спектрограм музики: модель (на основі Stable Diffusion) отримувала текстовий запит і генерувала спектрограму, яка потім перетворювалася на аудіо (Hadjeres G., Pachet F., Nielsen F., 2017). Таким чином, фактично було використано текст-до-зображення дифузійну модель, переналаштовану на простір аудіо. Цей підхід довів гнучкість дифузійних методів і відкрив шлях до використання напрацювань комп'ютерного зору в аудіо.

З 2023 року домінуючими моделями, що доступні широкому загалу користувачів, у тому числі безкоштовно, є моделі SUNO (suno.com) та UDIO (udio.com). На даний момент, оскільки ці моделі є комерційними, а не дослідницькими, наявна дуже обмежена кількість інформації про їх побудову та структуру фонкціонування. Порівняно із більш ранніми зразками, ці меделі демонструють якісно новий рівень аудіо генерації, а також надають можливість отримати результат

генерації як у вигляді мультитреку, із використанням технологій розділення джерел, так і у вигляді символьної генерації, тобто результат генерації конвертується у MIDI файли, які доступні для завантаження.

Висновки. Досліджено основні типи нейронних мереж глибокого навчання (ШІ) в контексті автономної генерації музичних композицій за текстовим запитом. Моделі символьної генерації демонструють значний прогрес у завданнях гармонізації, побудови логічних музичних послідовностей і стилізації під різні жанри. Застосування рекурентних мереж (LSTM) дало можливість моделювати локальні й довготривалі залежності, тоді як впровадження архітектури трансформерів із механізмом істотно збільшило гнучкість та ефективність роботи з музичними послідовностями. Завдяки цьому поліпшилися якість і структурна цілісність згенерованих композицій, зокрема й у жанрах, що передбачають поліфонічну багатошаровість. Використання GAN-моделей (на прикладі MuseGAN) підтвердило потенціал змагального навчання для синтезу складних музичних фрагментів із кількома доріжками. Водночас, попри позитивні слухацькі оцінки, більшість символьних моделей потребують додаткового етапу «озвучення» (рендерінгу) за допомогою семплерів, віртуальних інструментів тощо, що з одного боку збагачує можливості втручання у генерацію, залучаючи людину у такий творчий процес, а з іншого – обмежує безпосередню готовність підсумкового аудіопродукту. У сфері генерування аудіосигналів безпосередньо помітні успіхи дифузійних моделей, що дозволяють отримувати реалістичніші звучання та враховувати семантичні аспекти тексту й культурного контексту. Водночас певні обмеження (артефакти звучання, повтори, надмірна складність навчання) ускладнюють їх широке використання. Попри це, поява комерційних рішень підтверджує високу перспективність галузі. І попри те, що наразі професійний рівень деяких аспектів музичного виробництва із застосуванням таких моделей залишається недосяжним або надто ресурсоємним, загальна динаміка вказує на стрімке наближення систем генерації до реальних творчих і виробничих потреб.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., & Frank, C. MusicLM: Generating music from text. 2023. URL: https://arxiv.org/pdf/2301.11325 (дата звернення 1.04.2025).
2. Bazin, T., & Hadjeres, G. NONOTO: A model-agnostic web interface for interactive music composition by inpainting. *Proceedings of the 10th International Conference on Computational Creativity (ICCC 2019)*. URL: https://www.semanticscholar.org/reader/e3fa5116ac050a61682093df07c117718536a074 (дата звернення 1.04.2025).
3. Chen, Y., Huang, L., & Gou, T. Applications and advances of artificial intelligence in music generation: A review. URL: https://arxiv.org/pdf/2409.03715 (дата звернення 1.04.2025).
4. Civit, M., Civit-Masot, J., Cuadrado, F., & Escalona, M. J. A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. *Expert Systems with Applications*. Volume 209. 2022. URL: https://www.sciencedirect.com/science/article/pii/S0957417422013537 (дата звернення 1.04.2025).
5. Patrício da Silva. David Cope and experiments in musical intelligence. Spectrum Press, 2003. URL: https://eclass.uoa.gr/modules/document/file.php/MUSIC124/%CE%94%CE%B9%CE%B1%CE%BB%CE%AD%CE%BE%CE%B5%CE%B9%CF%82/da-silva-david-cope-and-emi.pdf (дата звернення 1.04.2025).
6. Dong, H.-W. *Generative AI for music and audio* (Doctoral dissertation). University of California, San Diego. 2024 URL: https://arxiv.org/pdf/2411.14627 (дата звернення 1.04.2025).
7. Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., & Yang, Y.-H.. MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. 2017 URL: https://www.researchgate.net/publication/328305824_MuseGAN_Multi-track_Sequential_Generative_Adversarial_Networks_for_Symbolic_Music_Generation_and_Accompaniment (дата звернення 1.04.2025).
8. Edwards, M. Algorithmic composition: Computational thinking in music. *Communications of the ACM*, 2011. 54(7), 58–67.
9. Sifei Li, Yuxin Zhang, Fan Tang, Chongyang Ma, Weiming Dong, Changsheng Xu. Music Style Transfer with Time-Varying Inversion of Diffusion Models. URL: https://arxiv.org/pdf/2402.13763 (дата звернення 1.04.2025).
10. Hadjeres G., Pachet F., Nielsen F. DeepBach: a steerable model for Bach chorales generation. *Proceedings of the 34th International Conference on Machine Learning* (PMLR 70). 2017. URL: https://proceedings.mlr.press/v70/hadjeres17a.html (дата звернення: 01.04.2025).
11. Hernandez-Olivan, C., Hernandez-Olivan, J., & Beltran, J. R. A survey on artificial intelligence for music generation: Agents, domains and perspectives. URL: https://arxiv.org/pdf/2210.13944 (дата звернення 1.04.2025).
12. Huang, Q., Park, D. S., Wang, T., Denk, T. I., Ly, A., Chen, N., Zhang, Z., Zhang, Z., Yu, J., Frank, C., Engel, J., Le, Q. V., Chan, W., Chen, Z., & Han, W. Noise2Music: Text-conditioned music generation with diffusion models. 2023. URL: https://arxiv.org/pdf/2302.03917 (дата звернення 1.04.2025).
13. Huang, Y.-S., & Yang, Y.-H. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)* (pp. 1180–1188). URL: https://arxiv.org/pdf/2002.00212 (дата звернення 1.04.2025).
14. Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brebisson, A., Bengio, Y., & Courville, A. MelGAN: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*. URL: https://arxiv.org/pdf/2106.07889 (дата звернення 1.04.2025).
15. Mathews, M. V. The digital computer as a musical instrument. *Science*, 142(3592), 1963. P. 553–557. URL: https://digitalmusicacademy.ru/sites/default/files/content/mathews-1963.pdf (дата звернення 4.04.2025).

16. Parmar, H. *Tradi-fusion refined: Evaluating and fine-tuning the Riffusion model for Irish traditional music* (Master's thesis). KTH Royal Institute of Technology, Stockholm, Sweden. 2024. URL: https://hdparmar.github.io/Tradifusion (дата звернення 01.04.2025 р.)

17. Schneider, F., Kamal, O., Jin, Z., & Schölkopf, B. Moûsai: Efficient text-to-music diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics.* 2024. P. 8050–8068. URL: https://aclanthology.org/2024.acl-long.437.pdf (дата звернення 1.04.2025).

18. Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., & Yu, D. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31. 2023. P. 1720–1733. URL: https://arxiv.org/pdf/2207.09983 (дата звернення 1.04.2025).

19. Zhang, N., Yan, J., & Briot, J.-P. Artificial intelligence techniques for pop music creation: A real music production perspective. Preprint submitted to *Information Fusion*, 2023. URL: https://hal.science/hal-04363458v1/document (дата звернення 01.04.2025 р.).

20. Zhao, Y., Yang, M., Lin, Y., Zhang, X., Shi, F., Wang, Z., Ding, J., & Ning. AI-enabled text-to-music generation: A comprehensive review of methods, frameworks, and future directions. *Electronics, 14* (6). 2025. P. 1197. URL: https://www.mdpi.com/2079-9292/14/6/1197 (дата звернення 1.04.2025).

REFERENCES

1. Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., & Frank, C. (2023). MusicLM: Generating music from text. URL: https://arxiv.org/pdf/2301.11325

2. Bazin, T., & Hadjeres, G. NONOTO: A model-agnostic web interface for interactive music composition by inpainting. (2019). *Proceedings of the 10th International Conference on Computational Creativity (ICCC 2019)*. URL: https://www.semanticscholar.org/reader/e3fa5116ac050a61682093df07c117718536a074

3. Chen, Y., Huang, L., & Gou, T. (2024). Applications and advances of artificial intelligence in music generation: A review. URL: https://arxiv.org/pdf/2409.03715

4. Civit, M., Civit-Masot, J., Cuadrado, F., & Escalona, M. J. (2022). A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. *Expert Systems with Applications*. Volume 209. URL: https://www.sciencedirect.com/science/article/pii/S0957417422013537

5. Patrício da Silva. (2003). David Cope and experiments in musical intelligence. Spectrum Press. URL: URL: https://eclass.uoa.gr/modules/document/file.php/MUSIC124/%CE%94%CE%B9%CE%B1%CE%BB%CE%AD%CE%BE%CE%B5%CE%B9%CF%82/da-silva-david-cope-and-emi.pdf

6. Dong, H.-W. (2024). *Generative AI for music and audio* (Doctoral dissertation). University of California, San Diego. URL: https://arxiv.org/pdf/2411.14627

7. Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., & Yang, Y.-H. (2017). MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. URL: https://www.researchgate.net/publication/328305824_MuseGAN_Multi-track_Sequential_Generative_Adversarial_Networks_for_Symbolic_Music_Generation_and_Accompaniment

8. Edwards, M. (2011). Algorithmic composition: Computational thinking in music. *Communications of the ACM*. 54(7).

9. Sifei Li, Yuxin Zhang, Fan Tang, Chongyang Ma, Weiming Dong, Changsheng Xu. (2024). Music Style Transfer with Time-Varying Inversion of Diffusion Models. URL: https://arxiv.org/pdf/2402.13763 (date of application 01.04.2025).

10. Hadjeres G., Pachet F., Nielsen F. DeepBach: a steerable model for Bach chorales generation. *Proceedings of the 34th International Conference on Machine Learning* (PMLR 70). 2017. URL: https://proceedings.mlr.press/v70/hadjeres17a.html (date of application 01.04.2025).

11. Hernandez-Olivan, C., Hernandez-Olivan, J., & Beltran, J. R. (2022). A survey on artificial intelligence for music generation: Agents, domains and perspectives. URL: https://arxiv.org/pdf/2210.13944 (date of application 01.04.2025).

12. Huang, Q., Park, D. S., Wang, T., Denk, T. I., Ly, A., Chen, N., Zhang, Z., Zhang, Z., Yu, J., Frank, C., Engel, J., Le, Q. V., Chan, W., Chen, Z., & Han, W. (2023). Noise2Music: Text-conditioned music generation with diffusion models. URL: https://arxiv.org/pdf/2302.03917 (date of application 01.04.2025).

13. Huang, Y.-S., & Yang, Y.-H. (2020). Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)* (pp. 1180–1188). https://doi.org/10.1145/3394171.3413671 (date of application 01.04.2025).

14. Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brebisson, A., Bengio, Y., & Courville, A. (2019). MelGAN: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*. URL: https://arxiv.org/pdf/2106.07889 (date of application 01.04.2025).

15. Mathews, M. V. (2009). The digital computer as a musical instrument. *Science*, 142(3592). URL: https://digitalmusicacademy.ru/sites/default/files/content/mathews-1963.pdf (date of application 04.04.2025).

16. Parmar, H. (2024). *Tradi-fusion refined: Evaluating and fine-tuning the Riffusion model for Irish traditional music* (Master's thesis). KTH Royal Institute of Technology, Stockholm, Sweden. URL: https://hdparmar.github.io/Tradifusion (date of application 01.04.2025).

17. Schneider, F., Kamal, O., Jin, Z., & Schölkopf, B. (2024). Moûsai: Efficient text-to-music diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics.* URL: https://aclanthology.org/2024.acl-long.437.pdf (date of application 01.04.2025).

18. Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., & Yu, D. (2023). Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31. URL: https://arxiv.org/pdf/2207.09983 (date of application 01.04.2025).

19. Zhang, N., Yan, J., & Briot, J.-P. (2023). Artificial intelligence techniques for pop music creation: A real music production perspective. Preprint submitted to *Information Fusion*. URL: https://hal.science/hal-04363458v1 (date of access: 01.04.2025).

20. Zhao, Y., Yang, M., Lin, Y., Zhang, X., Shi, F., Wang, Z., Ding, J., & Ning, H. (2025). AI-enabled text-to-music generation: A comprehensive review of methods, frameworks, and future directions. *Electronics, 14*(6), 1197. https://doi.org/10.3390/electronics14061197