DOI: https://doi.org/10.15276/ict.02.2025.05
УДК 004.4’24

Порівняльний аналіз точності класифікації ансамблевих та індивідуальних моделей на прикладі даних фейкового мовлення

Арсірій Олена Олександрівна[1)]
Д-р техніч. наук, професор каф. Інформаційних систем
ORCID: https://orcid.org/0000−0001−8130−9613; e.arsiriy@gmail.com. Scopus Author ID: 54419480900
Андронаті Олександр Кирилович[1)]
Аспірант каф. Інформаційних систем
ORCID: https://orcid.org/0009-0009-1794-5864; alex.andronati@gmail.com. Scopus Author ID: 58677655800
[1)] Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

АНОТАЦІЯ

У сучасних умовах швидкого розвитку технологій глибокого навчання, синтезоване мовлення на основі deepfake створює значні ризики для інформаційної безпеки, включаючи маніпуляції в медіа та кіберзлочини. Дослідження присвячене порівняльному аналізу точності класифікації ансамблевих та індивідуальних моделей для виявлення фейкового мовлення. Метою є підвищення ефективності детекції діпфейків шляхом розробки ансамблевих класифікаторів на основі stacking з стратегіями агрегації прогнозів (hard voting, soft voting та soft voting з нечітким ранжуванням Гомперца). Використано датасет Fake or Real. В якості базових моделей використано K-nearest neighbors, support vector machines, random forest, extreme gradient boosting, logistic regression, multilayer perceptron, convolutional neural networks та long short-term memory. З 657 ансамблів найкращий досяг accuracy 0,935 та F1-score 0,935, що на 3,9 % перевищує індивідуальні моделі. Результати підтверджують перевагу ансамблевих підходів у роботі з діпфейками.

Ключові слова: Deepfake; фейкове мовлення; ансамблева класифікація; машинне навчання; нейронні мережі; stacking; hard voting, soft voting

Актуальність. Швидкий розвиток технологій синтезу мовлення на основі глибокого навчання призвів до появи реалістичних аудіо deepfake-записів, які здатні точно імітувати голос людини. Такі технології активно використовуються у соціальних мережах, кіберзлочинності та медіа-маніпуляціях, створюючи суттєві ризики для інформаційної безпеки та цифрової довіри. Зважаючи на це, актуальною науковою проблемою є розробка методів автоматичного виявлення фейкового мовлення, які поєднують високу точність класифікації та стійкість до шуму.

Більшість сучасних методів детекції базуються на аналізі спектральних ознак (MFCC, STFT) у поєднанні з класичними або нейронними моделями. Проте, як показують дослідження [1–3], навіть глибокі архітектури згорткових нейронних мереж та мереж з довгою короткостроковою пам'яттю демонструють недостатньо високу точність на тестових вибірках, особливо коли дані містять шум або отримані з різних джерел. Це свідчить про обмежену узагальнювальну здатність індивідуальних моделей і підкреслює потребу у поєднанні декількох класифікаторів і дослідження ансамблевих стратегій класифікації, здатних підвищити стабільність результатів та знизити кількість помилкових класифікацій.

Метою дослідження є підвищення точності виявлення діпфейків шляхом розробки методу створення ансамблевих класифікаторів та порівняння ефективності класичних методів машинного навчання з ансамблевою класифікацією.

Завдання виявлення фейкового мовлення належить до класу бінарних задач класифікації, у яких необхідно віднести кожен аудіофайл до одного з двох класів: 0 – реальне мовлення, 1 – синтетичне (deepfake) мовлення.

Особливістю цієї задачі є поєднання часових, спектральних і статистичних ознак, що вимагає комплексного підходу до обробки даних і вибору моделей.

Для дослідження був обраний датасет Fake or Real (FoR) [4], який містить 17870 аудіофайлів для виявлення синтезованої мови (реальні та підроблені аудіофайли). Набір даних було розділено на 13 956 файлів у навчальному наборі та 3914 файлів у тестовому наборі.

Для максимально повного аналізу класифікації дипфейків набір даних був представлений у трьох видах, що дозволило застосувати різні підходи машинного навчання. По-перше, табличні спектральні характеристики були використані для класичних алгоритмів машинного навчання, забезпечуючи структурований аналіз числових ознак. По-друге, спектрограми як зображення були адаптовані для згорткових нейронних мереж, що ефективно враховує візуальні патерни в аудіоданих. По-третє, мел-частотні кестральні коефіцієнти (MFCC) як часові ряди були застосовані для мереж з довгою короткостроковою пам'яттю, дозволяючи моделювати послідовні залежності в сигналах.

Згідно з дослідженнями [5, 6, 7], найбільш перспективними спектральними характеристиками для класифікації аудіоданих є спектральний центроїд, спектральна рівномірність, спектральний контраст, спектральний спад, частота перетину нуля, MFCC.

MFCC – коефіцієнти, що виводяться з типу оберненого перетворення Фур'є. MFCC дозволяють краще представляти звук, оскільки частотні діапазони рівномірно розподілені по шкалі Мела, яка більш точно наближається до реакції людської слухової системи.

Мел – це одиниця висоти звуку, заснована на сприйнятті цього звуку нашими органами слуху. Оскільки частотна характеристика людського вуха не є лінійною, амплітуда не є цілком точним показником гучності звуку. Аналогічно, висота звуку, що сприймається людським слухом, не залежить лінійно від його частоти.

Ця залежність описується формулою:

$$m = 1125\ ln\ \left(1 + \frac{f}{700}\right), \tag{1}$$

де f – частота в герцах.

Графік, що показує залежність висоти звуку в мелах від частоти коливань, наведено на рис. 1.

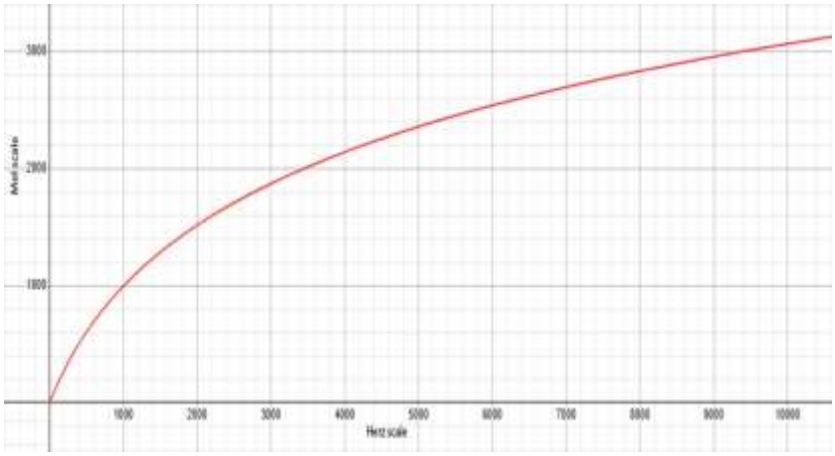

Рис. 1. **Залежність висоти звуку в мелах від частоти коливань**

Наступним кроком в підході до класифікації даних діпфейків є вибір набору з восьми основних класифікаторів, що охоплюють як класичні методи машинного навчання, так і глибокі нейронні мережі. Набір включає: K-найближчих сусідів (KNN), машини опорних векторів (SVM), логістичну регресію (LR), Random Forest, extreme gradient boosting (XGBoost), багатошаровий перцептрон (MLP), згорткову нейронну мережу (CNN), довгу короткострокову пам'ять (LSTM).

Цей вибір був обумовлений необхідністю поєднати моделі з різними сильними сторонами. Цей набір дозволяє адаптувати моделі до неоднорідних аудіоданих.

Розробка моделей розпочалася з визначення їх базової архітектури та параметрів.

У таблиці 1 наведено перелік основних гіперпараметрів для кожного алгоритму.

У випадках, коли було складно вручну вибрати оптимальні гіперпараметри для моделей, використовувалася технологія Grid Search Cross Validation (GSCV) з 5-кратною кросс-

валідацією на навчальних зразках, щоб уникнути перенавчання і забезпечити узагальнення. Для кожної моделі було встановлено діапазон параметрів (наприклад, для XGBoost сітка гіперпараметрів включала варіації швидкості навчання (0,01; 0,05; 0,1; 0,2; 0,3), максимальної глибини дерева (3, 5, 7, 10) та кількості дерев (50, 100, 200, 300), далі було обрано комбінації, що максимізували F1-score або accuracy на валідаційній вибірці.

Таблиця 1. **Гіперпараметри алгоритмів**

Алгоритм	Гіперпараметри
KNN	1. Кількість сусідів, які будуть використовуватися. 2. Метрика для обчислення відстані
LR	1. Максимальна кількість ітерацій
SVM	1. Тип ядра. 2. Параметр регуляризації C. 3. Ступінь функції ядра (якщо тип ядра є поліноміальним).
Random Forest	1. Кількість дерев 2. Максимальна глибина дерева. 3. Мінімальна кількість екземплярів, необхідних для наявності в кінцевому вузлі.
XGBoost	1. Кількість дерев 2. Максимальна глибина дерева. 3. Мінімальна сума ваги екземпляра, необхідна в дочірньому вузлі. 4. Швидкість навчання 5. Гамма – мінімальне зменшення втрат, необхідне для подальшого поділу кінцевого вузла дерева
MLP	1. Кількість прихованих шарів 2. Кількість нейронів у кожному шарі 3. Функція активації прихованих шарів 4. Відсоток у dropout-шарах 5. Оптимізатор
CNN	1. Кількість згорткових шарів 2. Кількість пулінг-шарів 3. Кількість повнозв'язних шарів 4. Кількість нейронів у них 5. Функції активації шарів 6. Оптимізатор
LSTM	1. Кількість шарів LSTM 2. Кількість нейронів у шарах LSTM 3. Кількість повнозв'язних шарів 4. Кількість нейронів у повнозв'язних шарах 5. Функції активації шарів 6. Функція рекурентної активації 7. Оптимізатор

Контроль перенавчання також було реалізовано шляхом моніторингу різниці в метриках між навчальними та тестовими метриками, для нейромережевих алгоритмів було використано early stopping.

Після отримання оптимальних гіперпараметрів усі алгоритми навчались на повній навчальній вибірці.

Наступним етапом запропонованого підходу є створення ансамблю класифікаторів. Існує багато різних типів ансамблювання, але для об'єднання різноманітних класифікаторів, які використовують різні види представлення даних, найкраще підходить тип ансамблювання stacking.

Для агрегації прогнозів у stacking використовуються такі підходи: hard voting, soft voting, та soft voting з нечітким ранжуванням Гомперца.

При використанні hard voting остаточне рішення ансамблю базується на більшості голосів моделей, кожна модель робить прогноз, і вибирається варіант з найбільшою кількістю голосів. Цей метод підходить для збалансованих класифікаторів з надійними прогнозами.

Розрахунок hard voting:

$$\hat{y} = \text{mode}(y_1, y_2, \ldots, y_n), \tag{2}$$

де $\hat{y}$ – остаточний прогнозований клас, визначений ансамблем;

y_i – прогнозований клас і-м індивідуальним класифікатором, де і варіюється від 1 до n;

n – загальна кількість індивідуальних класифікаторів в ансамблі.

У soft voting, кожен класифікатор присвоює класам ймовірності, а остаточне рішення визначається зваженим середнім цих ймовірностей.

Розрахунок soft voting:

$$p_k = (1/n) \Sigma\, p_{ik}, \qquad \hat{y} = \text{argmax}\, p_k, \tag{3}$$

де $\hat{y}$ – остаточний прогнозований клас, визначений ансамблем;

p_{ik} – ймовірність класу k, прогнозованого і-м класифікатором, де i – від 1 до n, а k – індекс класу;

n – загальна кількість окремих класифікаторів в ансамблі.

Soft voting з нечітким ранжуванням Гомперца дозволяє враховувати невизначеність і неоднорідність даних, що покращує точність і стабільність ансамблевого класифікатора при класифікації аудіоданих [8].

Розрахунок soft voting з використанням нечітких рангів (Gompertz):

$$p'_{i,k} = a \cdot exp(-b \cdot exp(-c \cdot p_{i,k})), \; p_k = (1/n) \Sigma\, p'_{i,k}, \qquad \hat{y} = \text{argmax}_k\, p_k, \tag{4}$$

де $\hat{y}$ – остаточний прогнозований клас, визначений ансамблем;

$p'_{i,k}$ – скоригована ймовірність класу k для і-го класифікатора після застосування функції Гомперца;

p_k – середня ймовірність класу k для всіх класифікаторів в ансамблі;

n – загальна кількість окремих класифікаторів в ансамблі;

a, b, c – параметри функції Гомперца, що використовуються для коригування ймовірностей, де a контролює верхню асимптоту, а b і c контролюють форму кривої.

В рамках запропонованої методології, були створені ансамблеві класифікатори, що складалися з різних комбінацій алгоритмів-складових.

З огляду на те, що є три типи агрегації прогнозів і вісім базових класифікаторів, отримуємо наступну кількість ансамблевих класифікаторів, тут C_n^k – кількість комбінацій від n до k:

$$C_n^k = 3\,(C_8^3 + C_8^4 + C_8^5 + C_8^6 + C_8^7 + C_8^8) = 657.$$

Незважаючи на побудову великої кількості ансамблів з різними комбінаціями базових класифікаторів, обчислювальна складність процесу залишається помірною завдяки простоті етапу голосування. Основні обчислювальні витрати – це навчання базових класифікаторів, яке виконується один раз. Це дозволяє генерувати багато ансамблів, змінюючи склад

класифікаторів, без значного збільшення загального обчислювального навантаження, оскільки додаткові витрати на голосування мінімальні в порівнянні з навчанням моделі.

Ефективність усіх ансамблів оцінювалася на тестових даних за допомогою ключових показників якості класифікації: accuracy та F1-score, що дало комплексне уявлення про ефективність моделі та дозволило вибрати найкращі конфігурації на основі максимальних значень accuracy та F1-score. Використання цих показників є критично важливим для об'єктивної оцінки, оскільки вони охоплюють різні аспекти ефективності моделі та допомагають уникнути упередженості, пов'язаної з характеристиками даних. F1-score є особливо цінним в завданні із дисбалансом класів, оскільки він забезпечує баланс між мінімізацією false positives (precision) та false negatives (recall). В результаті був обраний найкращий ансамбль на основі комбінованого критерію максимальної accuracy та F1-score.

З 657 розроблених ансамблевих класифікаторів найкращі результати показав ансамбль, що складався з KNN, Random Forest, CNN із soft voting. Були отримані такі показники:

– Accuracy = 0,935 (+3,9 % відносно найкращого класифікатора в ансамблі);

– F1 Score = 0,935 (+3,9 % відносно найкращого класифікатора в складі).

Повна інформація про порівняння найкращого ансамблю з окремими класифікаторами наведена в таблиці 2.

Таблиця 2. **Класифікаційні метрики**

Алгоритм	**Клаисфікаційні метрики**			
	Accuracy	***F1-score***	***Різниця Accuracy***	***Різниця F1-score***
Ансамбль	***0.935***	***0.935***		
SVM	0.816	0.813	11.9 %	12.2 %
MLP	0.852	0.827	8.3 %	10.8 %
XGBoost	0.792	0.789	14.3 %	14.6 %
Random Forest	0.828	0.827	10.7 %	10.8 %
KNN	0.896	0.896	3.9 %	3.9 %
CNN	0.894	0.903	4.1 %	3.2 %
LSTM	0.881	0.876	5.4 %	5.9 %
LR	0.816	0.813	11.9 %	12.2 %

У результаті проведеного дослідження було продемонстровано, що використання ансамблевих класифікаторів на основі stacking з різними стратегіями агрегації прогнозів (hard voting, soft voting та soft voting з нечітким ранжуванням Гомперца) значно підвищує точність виявлення фейкового мовлення порівняно з індивідуальними моделями. На датасеті Fake or Real, що містить 17 870 аудіофайлів, найкращий ансамбль, складений з KNN, Random Forest та CNN з soft voting, досяг показників accuracy 0,935 та F1-score 0,935, що на 3,9 % перевищує результати найкращої індивідуальної моделі в складі ансамблю. Це підтверджує ефективність поєднання класичних алгоритмів машинного навчання з глибокими нейронними мережами, адаптованими до різних представлень даних: табличних спектральних характеристик, спектрограм як зображень та MFCC як часових рядів.

Висновки. Запропонований підхід не лише покращує стабільність класифікації в умовах шуму та різноманітності джерел аудіоданих, але й знижує кількість помилкових класифікацій, що є критичним для задач інформаційної безпеки. Обчислювальна ефективність методу, зосереджена на одноразовому навчанні базових класифікаторів, дозволяє масштабувати його на більші датасети без значного зростання ресурсів.

Перспективи подальших досліджень включають інтеграцію додаткових спектральних ознак, тестування на інших датасетах з реальними deepfake-записами та вдосконалення нечітких методів ранжування для підвищення узагальнювальної здатності ансамблів. Отримані результати можуть бути застосовані в системах моніторингу медіа-контенту,

кібербезпеки та соціальних мереж для протидії маніпуляціям на основі синтезованого мовлення.

СПИСОК ЛІТЕРАТУРИ

1. Khanjani Z., Watson G., Janeja V.P. “Audio Deepfakes: A Survey”. *Frontiers in Big Data.* 2023; 6: 1001063. DOI: https://doi.org/10.3389/fdata.2022.1001063.

2. Zhang B., Cui H., Nguyen V., Whitty M. “Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead. Sensors”. 2025; 25 (7). DOI: https://doi.org/10.3390/s25071989.

3. Altuncu E., Franqueira V. N. L., Li S. “Deepfake: Definitions, Performance Metrics and Standards, Datasets, and a Meta-Review”. *Frontiers in Big Data*. 2024; 7: 1400024. DOI: https://doi.org/10.3389/fdata.2024.1400024.

4. “The Fake-or-Real (FoR) Dataset (deepfake audio)”. – URL: https://www.kaggle.com/datasets/mohammedabdeldayem/the-fake-or-real-dataset/data [Accessed: 21.08.2025].

5. McFee B., Raffel C., Liang D., et al. “librosa: Audio and music signal analysis in python.” I*n Proceedings of the 14th Python in Science Conference*. 2015. p. 18–25.

6. Thukroo I. A. Bashir R., Giri K. J. “A comparison of cepstral and spectral featuresusing recurrent neural network for spoken language identification”. *Comput. Artif. Intell*. 2024; 2 (1). DOI: https://doi.org/10.59400/cai.v2i1.440.

7. Sharma G., Umapathy K., Krishnan S. “Trends in audio signal feature extraction methods”. *Applied Acoustics*. 2020; 158: 107020. DOI: https://doi.org/10.1016/j.apacoust.2019.107020.

8. Andronati O., Antoshchuk S., Babilunha O., Arsirii O., Nikolenko A., Mikhalev K. “A method of constructing ensemble classifiers for recognizing audio data of various nature”. 2024. p. 758–761, https://www.scopus.com/authid/detail.uri?authorId=58677655800.
DOI: https://doi.org/10.1109/ACIT62333.2024.10712469.

DOI: https://doi.org/10.15276/ict.02.2025.05
UDC 004.4’24

Comparative analysis of the classification accuracy of ensemble and individual models using the example of fake speech data

Olena O. Arsirii[1)]
Doctor of Engineering Sciences, Professor, Head of the Department of Information Systems
ORCID: https://orcid.org/0000−0001−8130−9613; e.arsiriy@gmail.com. Scopus Author ID: 54419480900
Oleksandr K. Andronati[1)]
Postgraduate Student of the Department of Information Systems
ORCID: https://orcid.org/0009-0009-1794-5864; alex.andronati@gmail.com. Scopus Author ID: 58677655800
[1)] Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044, Ukraine

ABSTRACT

In today's environment of rapid development of deep learning technologies, synthesized speech based on deepfakes poses significant risks to information security, including media manipulation and cybercrime. The study is devoted to a comparative analysis of the accuracy of ensemble and individual models for detecting fake speech. The goal is to improve the effectiveness of deepfake detection by developing ensemble classifiers based on stacking with prediction aggregation strategies (hard voting, soft voting, and soft voting with Gompertz fuzzy ranking). The Fake or Real dataset was used. K-nearest neighbors, support vector machines, random forest, extreme gradient boosting, logistic regression, multilayer perceptron, convolutional neural networks, and long short-term memory were used as base models. Of the 657 ensembles, the best achieved an accuracy of 0.935 and an F1-score of 0.935, which is 3.9 % higher than individual models. The results confirm the advantage of ensemble approaches in working with deepfakes.

Keywords: Deepfake; fake speech; ensemble classification; machine learning; neural networks; stacking; hard voting, soft voting