

[DOI 10.28925/2663-4023.2026.35.1423](https://doi.org/10.28925/2663-4023.2026.35.1423)

УДК 004.056.53

Руда Христина Степанівна

доктор філософії, доцент кафедри захисту інформації
Національний університет «Львівська політехніка», Львів, Україна
ORCID: 0000-0001-8644-411X
khrystyna.s.ruda@lpnu.ua

Кос Ігор Іванович

студент кафедри захисту інформації
Національний університет «Львівська політехніка», Львів, Україна
ORCID: 0009-0001-4558-5036
ihor.kos.mkb.2025@lpnu.ua

Сабодашко Дмитро Володимирович

доктор філософії, старший викладач кафедри захисту інформації
Національний університет «Львівська політехніка», Львів, Україна
ORCID: 0000-0003-1675-0976
dmytro.v.sabodashko@lpnu.ua

Швед Марія Євгенівна

кандидат технічних наук, доцент кафедри захисту інформації
Національний університет «Львівська політехніка», Львів, Україна
ORCID: 0000-0003-0428-7777
mariia.y.shved@lpnu.ua

Глушенко Павло Костянтинович

доктор філософії, асистент кафедри захисту інформації
Національний університет «Львівська політехніка», Львів, Україна
ORCID: 0000-0002-1262-5484
pavlo.k.hlushchenko@lpnu.ua

ВИЯВЛЕННЯ КЛОНОВАНОГО ГОЛОСУ КЛАСИЧНИМИ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Анотація. Стрімкий розвиток нейронних технологій синтезу мовлення з тексту (TTS) та перетворення голосу (VC) дав змогу створювати надзвичайно реалістичне синтезоване мовлення, що суттєво підвищило ризик атак підроблення голосу та атак соціальної інженерії. У роботі досліджено виявлення синтезованого мовлення класичними методами машинного навчання на основі акустичних ознак, сформованих вручну. Для експериментальної оцінки сформовано власний набір даних обсягом 3600 записів: 400 записів справжнього мовлення 20 мовців і 3200 записів, синтезованих чотирма сучасними системами клонування голосу та синтезу мовлення — RVC, XTTS, ElevenLabs і Tortoise (співвідношення класів 1 : 8). Кожен запис описано вектором зі 135 кепстральних, спектральних і часових ознак, який подається на вхід семи класичних класифікаторів, зокрема машин опорних векторів, ансамблевих методів і методу k найближчих сусідів. Найкращий результат отримано для SVM з ядром RBF: збалансована точність становить 91,1 %, при цьому модель правильно приймає 85 % справжніх записів і виявляє 97,2 % синтезованих. Аналіз за системами синтезу показав, що основним джерелом помилок є хибне відхилення справжнього мовлення: для чотирьох із семи моделей частка правильно прийнятих справжніх записів становить лише 27–64 %, тоді як синтезоване мовлення виявляється з точністю не нижче 85 %. Мовлення комерційної системи ElevenLabs виявилось найскладнішим для виявлення лінійними моделями. Отримані результати є базовими показниками для сценарію, у якому мовці та системи синтезу спільні для навчальної і тестової підмножин, і можуть слугувати орієнтиром для порівняння з детекторами на основі глибокого навчання.



Ключові слова: аудіодипфейк; виявлення синтезованого мовлення; клонування голосу; протидія підробленню голосу; класичне машинне навчання; машина опорних векторів; верифікація мовця.

ВСТУП

Останні досягнення в галузі синтезу мовлення з тексту (text-to-speech, TTS) та перетворення голосу (voice conversion, VC) дали змогу створювати надзвичайно реалістичне синтезоване мовлення, яке дедалі складніше відрізнити від справжнього людського голосу. Сучасні нейронні архітектури, зокрема моделі на основі трансформерів, дифузійні та авторегресійні моделі, здатні точно відтворювати індивідуальні особливості мовця, просодію та акустичні характеристики голосу [1]. Ці технології мають легітимні застосування, наприклад голосові помічники та засоби доступності, проте зловживання ними створює значні ризики для безпеки та суспільства.

Синтезоване мовлення можна використовувати для обходу систем голосової автентифікації, проведення голосового фішингу (вішингу) та поширення дезінформації через сфабриковані аудіозаписи. У криміналістичному та юридичному контекстах наявність переконливого синтезованого мовлення ускладнює встановлення автентичності аудіозаписів і оцінювання їх доказової сили [2]. Тому надійне виявлення синтезованого та маніпульованого мовлення стало однією з ключових задач біометричної безпеки та аудіокриміналістики.

Ранні роботи з виявлення синтезованого мовлення та протидії підробленню голосу були зосереджені переважно на захисті систем автоматичної верифікації мовця (automatic speaker verification, ASV) за допомогою акустичних ознак, сформованих вручну, і класичних класифікаторів [3]. У пізніших дослідженнях запропоновано детектори на основі глибокого навчання — згорткові, трансформерні та наскрізні (end-to-end) архітектури, які демонструють високу ефективність у контрольованих умовах оцінювання [4–6].

Постановка проблеми. Попри суттєвий прогрес у розробленні детекторів синтезованого мовлення, низка проблем залишається невирешеною: узагальнення на методи синтезу, не представлені в навчальних даних, стійкість до зміни якості аудіо та інтерпретованість рішень детектора [7]. Водночас нові загальнодоступні та комерційні системи клонування голосу з'являються швидше, ніж оновлюються стандартні оцінювальні корпуси, тож для таких систем бракує навіть базових оцінок того, наскільки легко виявити згенероване ними мовлення. У цьому контексті класичні методи машинного навчання залишаються актуальними як обчислювально ефективні базові моделі: хоча за точністю їх перевершують глибокі моделі, вони дають змогу контрольовано аналізувати поведінку класифікаторів за різних умов. Тому встановлення класичних базових показників для сучасних систем синтезу є важливим для коректної оцінки переваг складніших архітектур.

У статті досліджено виявлення мовлення, синтезованого кількома сучасними системами TTS та клонування голосу. Систематичну оцінку виконано для набору класичних класифікаторів машинного навчання, що працюють з акустичними ознаками, сформованими вручну. Дослідження зосереджене на порівнянні ефективності виявлення для різних систем синтезу та на аналізі поведінки класифікаторів за стандартними метриками, прийнятими в дослідженнях з біометрії та протидії підробленню голосу.

На відміну від панівної тенденції до дедалі складніших нейронних архітектур, ця робота свідомо зосереджена на класичних методах машинного навчання. Такі методи залишаються актуальними в чутливих до безпеки середовищах з обмеженими обчислювальними ресурсами завдяки низьким обчислювальним витратам і простоті розгортання. Класичні базові показники потрібні для змістовного порівняння з детекторами на основі глибокого навчання та для розуміння того, які артефакти синтезу можна виявити без наскрізного навчання представлень.

Наукова новизна та основні результати роботи такі:

1. сформовано набір даних, що містить справжнє мовлення та мовлення, синтезоване чотирма сучасними системами TTS і перетворення голосу, зокрема комерційною системою ElevenLabs, яка не представлена в стандартних корпусах ASVspoof;
2. для цього поєднання систем синтезу отримано порівняльну оцінку семи класичних класифікаторів на єдиному наборі зі 135 акустичних ознак за збалансованою точністю та точністю окремо для кожного класу;
3. виконано аналіз ефективності виявлення окремо для кожної системи синтезу та кожного класу, який показав, що основним джерелом помилок класичних детекторів є хибне відхилення справжнього мовлення, а не пропуск синтезованого;
4. отримано базові показники, придатні для подальшого порівняння з детекторами підробленого мовлення на основі глибокого навчання.



Аналіз останніх досліджень і публікацій. Стрімкий розвиток нейронних технологій TTS і VC суттєво розширив поверхню атаки для систем верифікації мовця, сервісів з голосовим керуванням і сценаріїв соціальної інженерії. Сучасні платформи синтезу на основі трансформерів, дифузійних моделей і нейронних вокодерів генерують мовлення високої якості з високою схожістю з цільовим мовцем, тому ні слухач, ні багато традиційних засобів захисту на основі обробки сигналів уже не забезпечують надійного розрізнення. Як зазначено в оглядовій роботі [1], основна складність виявлення аудіодипфейків змістилася від ідентифікації очевидних артефактів синтезу до узагальнення на невідомі методи генерації, акустичні середовища та канали передавання.

Для систематичної оцінки та порівняння детекторів наукова спільнота запровадила низку стандартизованих наборів даних і змагань. Основною платформою оцінювання засобів протидії підробленню в системах ASV стала серія змагань ASVspoof. Ранні її етапи були зосереджені на атаках логічного доступу на основі TTS і VC та на атаках повторного відтворення [8], тоді як пізніші етапи наголошують на реалістичних каналах передавання, стисненні та виявленні дипфейків [9]. Паралельно змагання з виявлення аудіодипфейків ADD роблять акцент на міждоменному узагальненні та виявленні сучасних методів нейронного синтезу в необмежених умовах [10]. Звіти за результатами цих змагань стабільно фіксують значний розрив між ефективністю детекторів на даних із того самого домену та поза ним, що свідчить про обмежену надійність багатьох наявних детекторів.

Ранні підходи до виявлення аудіодипфейків спиралися переважно на акустичні ознаки, сформовані вручну, — MFCC, CQCC, LFCC, фазові дескриптори — у поєднанні з класичними класифікаторами, зокрема моделями гаусових сумішей (GMM) і машинами опорних векторів [3, 8]. Ці методи обчислювально ефективні, але часто спираються на артефакти, специфічні для конкретного генератора, та на систематичні зміщення наборів даних. Як наслідок, їх ефективність істотно погіршується на невідомих методах синтезу та за реальних спотворень сигналу [7]. Водночас показано, що індивідуалізоване моделювання суттєво впливає на поведінку мовленнєвих систем, що підкреслює роль характеристик, специфічних для мовця, як у системах розпізнавання мовлення, так і в застосуваннях, чутливих до безпеки [11].

Сучасні найефективніші детектори базуються на глибокому навчанні, причому в новітній літературі переважають згорткові та трансформерні архітектури. Поширеним підходом є застосування глибоких моделей до часово-частотних представлень (мел-спектрограм, CQT, карт MFCC) з використанням згорткових мереж, зокрема LCNN [4], залишкових мереж і візуальних трансформерів, адаптованих для аналізу аудіо. Наскрізні підходи, як-от RawNet2 [5] і AASIST [6], а також архітектури з параметризованими фільтрами на кшталт SincNet [12] працюють безпосередньо з необробленою хвильовою формою. Глибокі моделі значно перевершують класичні базові моделі в контрольованих умовах, проте мають нижчу інтерпретованість і підвищену чутливість до особливостей конкретного набору даних. Окрім автономних детекторів, у системах контролю доступу успішно застосовуються інтегровані біометричні рішення на основі векторних представлень мовця (speaker embeddings), зокрема для виявлення сторонніх осіб [13].

Останнім часом перспективним напрямом підвищення надійності детекторів аудіодипфейків стало самокероване навчання (self-supervised learning, SSL). Моделі представлення мовлення, попередньо навчені на великих нерозмічених корпусах, використовують як фіксовані екстрактори ознак або як базові мережі, що донавчаються для виявлення підроблення. Зокрема, поєднання моделі wav2vec 2.0 з аугментацією даних суттєво покращило переносимість детекторів між наборами даних і типами атак [14]. Водночас в оглядових роботах наголошується, що детектори на основі SSL також схильні до перенавчання і можуть покладатися на хибні кореляції, якщо протоколи оцінювання недостатньо суворі [1]. Крім того, показано, що масштабованість та архітектура нейронних систем голосової біометрії суттєво впливають на надійність автентифікації [15], що зумовлює потребу в додаткових механізмах виявлення підроблення.

Оскільки багато практичних атак спрямовано саме на системи верифікації мовця, значна частина досліджень зосереджена на інтеграції засобів протидії підробленню безпосередньо в ланцюг обробки ASV, а не на розгляді виявлення як окремої задачі класифікації. Результати змагань ASVspoof показують, що практична ефективність залежить не лише від точності детектора, а й від узгодження його порога з порогом прийняття рішень системи ASV: надто агресивне відхилення підробленого мовлення призводить до неприйнятної частки хибних відхилень легітимних користувачів [8, 9]. Це зумовило розроблення метрик оцінювання, що враховують умови розгортання, та стратегій спільної оптимізації. Сучасні огляди підтверджують, що надійність і узагальнення залишаються невирішеними проблемами попри швидкий методологічний прогрес [1].

Виділення невирішених раніше частин загальної проблеми. Отже, попри значний прогрес у виявленні аудіодипфейків, узагальнення, надійність і застосовність у реальних умовах залишаються невирішеними проблемами. Зокрема: (1) стандартні оцінювальні корпуси не охоплюють новітніх загальнодоступних і комерційних систем клонування голосу, таких як RVC, XTTS та ElevenLabs; (2) для

цих систем бракує відтворюваних базових показників класичних детекторів, з якими можна було б порівнювати складніші моделі; (3) у більшості робіт ефективність подається узагальнено, без аналізу помилок окремо за класами та за системами синтезу, хоча для практичного розгортання саме частка хибних відхилень справжнього мовлення визначає зручність системи для легітимних користувачів.

Мета статті. Метою статті є побудова та експериментальна оцінка базових моделей класичного машинного навчання для виявлення мовлення, синтезованого сучасними системами клонування голосу, з порівнянням їх ефективності за стандартними метриками та аналізом придатності до використання в системах безпеки та аудіокриміналістики. Для досягнення мети поставлено такі завдання: сформувати набір даних справжнього та синтезованого мовлення; обґрунтувати набір акустичних ознак і класифікаторів; оцінити ефективність класифікаторів за збалансованою точністю та за точністю окремо для кожного класу й кожної системи синтезу; визначити обмеження класичних детекторів і напрями подальших досліджень.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Опис набору даних. Експерименти проведено на власному наборі даних обсягом 3600 аудіозаписів, що містить як справжнє, так і синтезоване мовлення. Набір призначений для оцінювання виявлення аудіодипфейків у контрольованих умовах для кількох сучасних систем синтезу мовлення. Структуру набору даних показано на рис. 1.

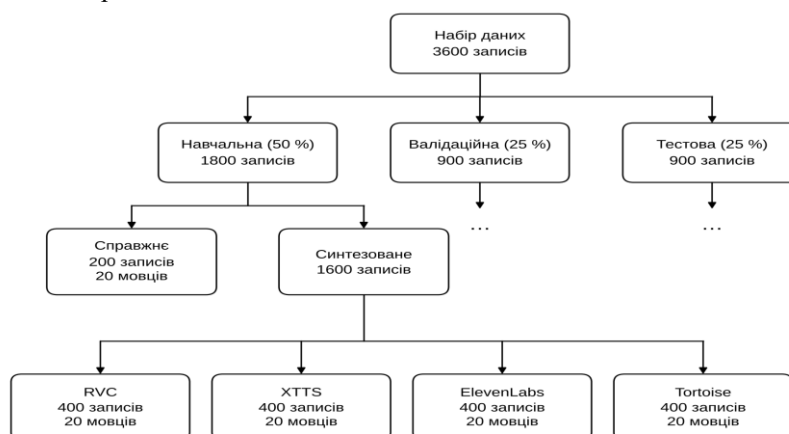


Рис. 1. Структура набору даних справжнього та синтезованого мовлення (наведено навчальну підмножину; валідаційна й тестова мають таку саму структуру)

Набір даних містить 400 записів справжнього мовлення 20 мовців (по 20 записів на мовця) та 3200 синтезованих записів — по 800 для кожної з чотирьох систем клонування голосу та синтезу мовлення: RVC, XTTS, ElevenLabs і Tortoise. Отже, набір є незбалансованим за класами: співвідношення справжніх і синтезованих записів становить 1 : 8, тоді як чотири системи синтезу представлені рівномірно. Усі записи — англійською мовою. Справжні записи — фрагменти аудіокниг волонтерського проєкту LibriVox, які перебувають у суспільному надбанні (public domain); записи отримано з платформи YouTube, тривалість кожного становить 1–3 хв. Синтезовані записи згенеровано для тих самих 20 мовців на основі їх справжнього мовлення, причому обсяг еталонного аудіо визначався вимогами кожної системи синтезу, а тексти для синтезу відрізнялися від текстів справжніх записів. Хоча вихідні записи перебувають у суспільному надбанні, окрема згода читців на клонування голосу не отримувалася, тому синтезовані записи використовуються виключно в дослідницьких цілях і не поширюються.

Набір даних розділено на навчальну (50 %, 1800 записів), валідаційну (25 %, 900 записів) і тестову (25 %, 900 записів) підмножини за допомогою стратифікованої процедури, що зберігає пропорції (i) класів (справжнє / синтезоване мовлення), (ii) систем синтезу та (iii) мовців. Отже, кожна підмножина містить один справжній запис на вісім синтезованих; зокрема, тестова підмножина містить 100 справжніх і 800 синтезованих записів (по 200 для кожної системи синтезу). Записи кожного мовця присутні в усіх трьох підмножинах; наслідки такого протоколу для інтерпретації результатів розглянуто в підрозділі «Обмеження дослідження».

Кожен запис має бінарну мітку (справжній або синтезований). Для синтезованих записів додатково зберігається назва системи синтезу, що дає змогу аналізувати ефективність виявлення окремо для кожного генератора.

Сценарій застосування детектора. Можливий сценарій застосування досліджуваного детектора в системі контролю доступу показано на рис. 2. Ця архітектура є концептуальною: у статті експериментально оцінюється лише модуль виявлення синтезованого мовлення, тоді як систему верифікації мовця не реалізовано й не оцінено. Детектор поєднується зі стандартною системою верифікації мовця (speaker verification, SV) за принципом багаторівневого захисту, щоб підвищити стійкість до атак з використанням аудіодипфейків і клонування голосу.

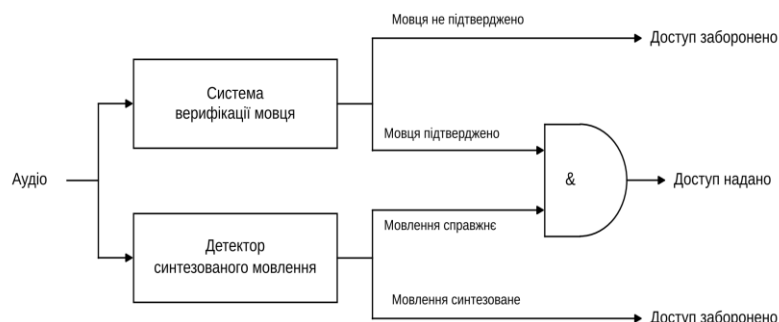


Рис. 2. Концептуальна архітектура системи контролю доступу з детектором синтезованого мовлення

Вхідне аудіовисловлювання може належати як справжньому мовцеві, так і синтетичному джерелу — системі синтезу мовлення з тексту або перетворення голосу. Сигнал паралельно надходить до двох незалежних модулів: детектора синтезованого мовлення, який аналізує акустичні характеристики сигналу й ухвалює рішення «справжнє» або «синтезоване», та системи верифікації мовця, яка порівнює векторне представлення мовця з еталоном, збереженим під час реєстрації, за допомогою міри подібності та порога прийняття рішення.

Рішення про надання доступу ухвалюється за логічною функцією «І»: доступ надається лише тоді, коли система верифікації підтвердила особу мовця і детектор класифікував мовлення як справжнє. Якщо хоча б одна з умов не виконується, запит на доступ відхиляється. Таким чином, навіть високоякісний голосовий клон, що успішно пройшов верифікацію мовця, буде відхилено, якщо детектор розпізнав його як синтезований.

Альтернативою є послідовний (каскадний) варіант, у якому до системи верифікації передаються лише сигнали, класифіковані детектором як справжні; з погляду рішення про доступ він еквівалентний паралельному, але зменшує обчислювальне навантаження на систему верифікації. В обох варіантах кожне хибне відхилення справжнього мовлення детектором означає відмову в доступі легітимному користувачеві, тому саме цей показник детально аналізується в розділі результатів.

Така модульна конструкція має кілька переваг. По-перше, детектор синтезованого мовлення та систему верифікації мовця можна розробляти, навчати й оцінювати незалежно. По-друге, детектор можна оновлювати або замінювати без модифікації основної біометричної системи. По-третє, архітектура відповідає реальним конвеєрам автентифікації, де засоби протидії підробленню зазвичай розгортаються як окремий захисний шар, а не як невід’ємна частина моделі верифікації.

Конвеєр виявлення синтезованого мовлення. Досліджуваний алгоритм виявлення, показаний на рис. 3, складається з трьох основних етапів: попередньої обробки аудіо, виокремлення ознак і класифікації.

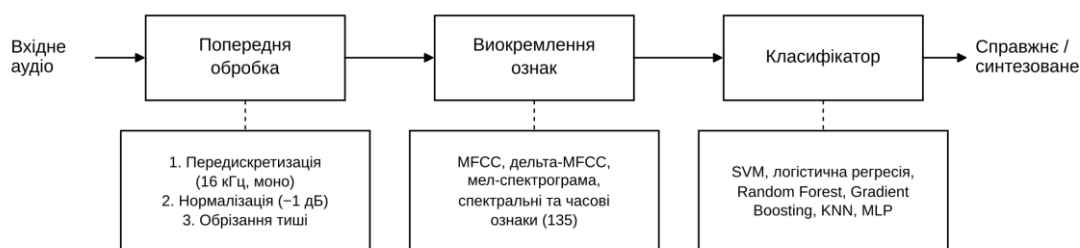


Рис. 3. Класичний алгоритм виявлення синтезованого мовлення



Параметри попередньої обробки (частота дискретизації, нормалізація, обрізання тиші) наведено в розділі «Методика дослідження».

Після попередньої обробки кожен аудіосигнал перетворюється на вектор ознак фіксованої розмірності, сформованих вручну. Ці ознаки покликані відобразити артефакти, які зазвичай вносять системи синтезу мовлення та перетворення голосу: неприродні спектральні обвідні, аномальну часову динаміку та невідповідності в розподілі енергії. Вибір частини ознак зумовлений відомими артефактами нейронних вокодерів, характерними для синтезованого мовлення [16].

Отримані вектори ознак подаються на вхід набору класичних класифікаторів машинного навчання, кожен з яких навчається розрізняти справжнє та синтезоване мовлення. Усі класифікатори використовують одне й те саме представлення ознак, однаковий поділ даних і однакову попередню обробку, що забезпечує порівнянність результатів.

Кожна модель ухвалює власне рішення незалежно від інших; результати моделей оцінюються окремо, без ансамблювання, що дає змогу об'єктивно порівняти різні класи методів.

Системи синтезу мовлення. У дослідженні класичні класифікатори машинного навчання розглядаються як базові засоби виявлення синтезованого мовлення. Вони реалізують традиційну двоетапну схему — виокремлення ознак, сформованих вручну, з подальшою класифікацією з учителем, — яка широко застосовувалася в попередніх дослідженнях з протидії підробленню голосу та аудіокриміналістики [3, 8].

Синтезоване мовлення в наборі даних згенеровано чотирма сучасними системами синтезу мовлення з тексту та перетворення голосу, що представляють різні парадигми синтезу (табл. 1).

Таблиця 1

Огляд систем синтезу мовлення, використаних у наборі даних

Система	Тип	Опис
ElevenLabs	Нейронна TTS (комерційна)	Комерційний хмарний сервіс синтезу мовлення з клонуванням голосу [17]
RVC	Перетворення голосу	Перетворення голосу з пошуком (retrieval) найближчих фрагментів ознак і декодером на основі VITS [18, 19]
Tortoise-TTS	Авторегресійна TTS	Багатомовцева система синтезу з клонуванням голосу: авторегресійний трансформер і дифузійний декодер [20]
XTTS	Авторегресійна TTS	Багатомовна система синтезу з клонуванням голосу за коротким зразком (zero-shot) [21]

Кожен аудіосигнал перетворюється на вектор зі 135 ознак, що описують кепстральні, спектральні та часові характеристики мовлення. Склад ознак і їх розмірність наведено в табл. 2.

Таблиця 2

Категорії акустичних ознак, сформованих вручну, та їх розмірність

Категорія ознак	Опис	Кількість
MFCC	13 коефіцієнтів MFCC; для кожного — середнє, стандартне відхилення, мінімум і максимум за кадрами	52
Дельта-MFCC	Похідні першого порядку 13 коефіцієнтів MFCC; зі статистичними дескрипторами	26
Мел-спектрограма	Глобальні та посмугові статистичні дескриптори мел-спектрограми	44
Спектральні	Статистичні дескриптори спектрального центроїда, ширини смуги, контрасту та спаду (rolloff)	8
Часові	Статистичні дескриптори частоти переходів через нуль (ZCR) і середньоквадратичної енергії (RMS)	5
Усього		135



Ознаки подаються на вхід семи класичних класифікаторів — лінійних і нелінійних. Класифікатори та їх основні параметри наведено в табл. 3.

Таблиця 3

Класифікатори та їх гіперпараметри

Класифікатор	Параметри	Опис
Random Forest	n_estimators = 100, max_depth = 20	Ансамбль дерев рішень
SVM (RBF)	kernel = rbf, class_weight = balanced, C — пошук по сітці	Нелінійна машина опорних векторів
SVM (Linear)	kernel = linear, class_weight = balanced, C — пошук по сітці	Лінійна машина опорних векторів
Gradient Boosting	n_estimators = 100, max_depth = 5	Модель послідовного бустингу
Logistic Regression	max_iter = 1000, class_weight = balanced, C — пошук по сітці	Лінійний імовірнісний класифікатор
K-Nearest Neighbors	n_neighbors = 5	Класифікатор на основі відстані
Multilayer Perceptron	hidden_layer_sizes = (256, 128), alpha — пошук по сітці	Багатошарова нейронна мережа прямого поширення

Для ансамблевих методів і KNN використано фіксовані значення гіперпараметрів, наведені в табл. 3; для SVM, логістичної регресії та MLP параметр регуляризації підбирався на валідаційній підмножині (див. розділ «Методика дослідження»).

Класичні методи машинного навчання мають низку практичних переваг — низькі обчислювальні витрати, швидке навчання та простоту розгортання. Водночас їх ефективність значною мірою залежить від вибору ознак, і вони, як правило, чутливі до зміни домену, зокрема до появи невідомих методів синтезу [7].

МЕТОДИКА ДОСЛІДЖЕННЯ

Виявлення синтезованого мовлення сформульовано як задачу бінарної класифікації. Позитивним класом вважається синтезоване мовлення, негативним — справжнє. Для вхідного аудіосигналу $x \in \mathbb{R}^{(N)}$, де N — кількість відліків, спочатку обчислюється вектор ознак $\phi(x) \in \mathbb{R}^{135}$, після чого класифікатор $f: \mathbb{R}^{135} \rightarrow [0, 1]$ оцінює ступінь належності сигналу до синтезованого класу. Бінарне рішення ухвалюється порівнянням значення $f(\phi(x))$ з порогом τ .

Попередня обробка аудіо. Усі аудіозаписи приведено до єдиної частоти дискретизації 16 кГц (моно). Амплітуду нормалізовано до піка -1 дБ для усунення відмінностей у рівні гучності між записами. Початкові та кінцеві ділянки тиші видалено пороговим алгоритмом на основі енергії з порогом 20 дБ нижче піка сигналу. Попередню обробку виконано засобами бібліотек librosa (версія 0.10.x) та NumPy.

Виокремлення ознак. З кожного попередньо обробленого сигналу обчислюється вектор зі 135 ознак. Аналіз виконується з вікном Ганна тривалістю 25 мс (400 відліків за 16 кГц) і кроком 10 мс (160 відліків). Набір ознак включає: 13 мел-частотних кепстральних коефіцієнтів (MFCC), обчислених на 40 мел-фільтрах, для кожного з яких обчислюються середнє, стандартне відхилення, мінімум і максимум за кадрами (52 ознаки); статистичні дескриптори дельта-MFCC першого порядку (26 ознак); статистичні дескриптори мел-спектрограми (44 ознаки); статистичні дескриптори спектральних ознак — центроїда, ширини смуги, контрасту і спаду (8 ознак); статистичні дескриптори часових ознак — частоти переходів через нуль і середньоквадратичної енергії (5 ознак). Усі ознаки стандартизовано до нульового середнього та одиничної дисперсії за статистиками навчальної підмножини (StandardScaler).

Розподіл даних. Набір даних (3600 записів) розділено стратифіковано за трьома критеріями — класом (справжнє / синтезоване мовлення), системою синтезу та мовцем — у пропорції 50 / 25 / 25 % на навчальну (1800 записів), валідаційну (900) і тестову (900) підмножини. Співвідношення класів 1 : 8 зберігається в кожній підмножині; тестова підмножина містить 100 справжніх і 800 синтезованих записів.



Класифікатори та підбір гіперпараметрів. Оцінено сім класичних класифікаторів: SVM з ядром RBF, SVM з лінійним ядром, логістичну регресію, випадковий ліс (Random Forest; 100 дерев, максимальна глибина 20), градієнтний бустинг (Gradient Boosting; 100 ітерацій, максимальна глибина 5), метод k найближчих сусідів (KNN, $k = 5$) і багатошаровий перцептрон (MLP; два приховані шари по 256 і 128 нейронів). Для SVM, логістичної регресії та MLP параметр регуляризації підбирався пошуком по сітці на валідаційній підмножині. Для обох варіантів SVM і логістичної регресії застосовано зважування класів ($class_weight = balanced$) для часткової компенсації дисбалансу; для ансамблевих методів, KNN і MLP зважування не застосовувалося. Усі класифікатори реалізовано засобами бібліотеки scikit-learn (версія 1.3.x).

Метрики оцінювання. Рішення класифікаторів отримано за порога $\tau = 0,5$; ймовірнісні оцінки для SVM отримано калібруванням Платта (параметр $probability = True$ у scikit-learn). Основною метрикою є збалансована точність (balanced accuracy) — середнє арифметичне часток правильно класифікованих справжніх і синтезованих записів; на відміну від звичайної точності, вона не залежить від співвідношення класів. Додатково наводяться точність (accuracy), влучність (precision), повнота (recall) та F1-міра для синтезованого (позитивного) класу, а також частка правильно класифікованих записів окремо для кожного класу та кожної системи синтезу. Через дисбаланс класів точність, влучність і F1-міру слід інтерпретувати обережно: тривіальний класифікатор, що відносить усі записи до синтезованого класу, отримав би на тестовій підмножині точність 88,9 %, влучність 88,9 %, повноту 100 % і F1-міру 94,1 %, але збалансовану точність лише 50 %. Усі метрики обчислено на тестовій підмножині, яка не використовувалася ні для навчання, ні для підбору гіперпараметрів.

Програмне середовище. Експерименти виконано в середовищі Python 3.10 з бібліотеками librosa, scikit-learn і NumPy. Навчання всіх моделей разом із підбором гіперпараметрів триває менше ніж 5 хвилин на CPU (Intel Core i7, 16 ГБ ОЗП), що підтверджує обчислювальну ефективність класичних підходів.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

У розділі наведено результати експериментальної оцінки класичних класифікаторів машинного навчання для виявлення синтезованого мовлення. Зведені метрики на тестовій підмножині подано в табл. 4.

Таблиця 4

Ефективність класичних класифікаторів на тестовій підмножині (позитивний клас — синтезоване мовлення)

Класифікатор	Точність, %	Влучність, %	Повнота, %	F1, %	Збалансована точність, %
Logistic Regression	90,0	98,1	90,5	94,1	88,2
SVM RBF	95,9	98,1	97,2	97,7	91,1
SVM Linear	90,0	98,0	90,6	94,2	87,8
Random Forest	91,6	91,6	99,6	95,4	63,3
Gradient Boosting	92,9	93,3	99,1	96,1	71,1
KNN	95,3	95,7	99,2	97,4	81,6
MLP	94,1	94,9	98,6	96,8	78,3

Значення метрик у табл. 4 обчислено з часток правильно класифікованих тестових записів кожного класу та кожної системи синтезу (рис. 4) за порога $\tau = 0,5$. Точність (90,0–95,9 %) і F1-міра (94,1–97,7 %) усіх моделей лише помірно перевищують рівень тривіального класифікатора (88,9 % і 94,1 % відповідно), а F1-міра лінійних моделей практично дорівнює йому. Натомість збалансована точність, що не залежить від дисбалансу класів, варіює значно ширше — від 63,3 до 91,1 % — і краще відображає реальні відмінності між моделями.

Найкращі результати демонструє SVM з ядром RBF: він має найвищі збалансовану точність (91,1 %), точність (95,9 %) і F1-міру (97,7 %), поєднуючи високу влучність (98,1 %) з високою повнотою



(97,2 %). Це вказує на те, що нелінійна межа рішення є ефективною для розмежування справжнього та синтезованого мовлення у використаному просторі ознак. Водночас справжній клас у тестовій підмножині представлений лише 100 записами: одна помилка змінює частку прийнятих справжніх записів на 1 п. п., а збалансовану точність — на 0,5 п. п., тому різниці між моделями в 1–2 п. п. не слід вважати статистично значущими.

Класифікатор KNN має другі за величиною точність (95,3 %) і F1-міру (97,4 %), проте його збалансована точність (81,6 %) помітно нижча, ніж у SVM з ядром RBF і лінійних моделей. Висока повнота (99,2 %) поєднується з нижчою влучністю (95,7 %), тобто KNN надійно виявляє синтезоване мовлення, але хибно відхиляє близько третини справжніх записів (див. рис. 4).

Лінійні моделі — логістична регресія та лінійний SVM — мають найнижчу повноту (90,5–90,6 %), але високу влучність (98,0–98,1 %, на рівні SVM з ядром RBF). За збалансованою точністю (88,2 % і 87,8 %) вони посідають друге й третє місця після SVM з ядром RBF: лінійні моделі частіше пропускають синтезоване мовлення, проте рідше хибно відхиляють справжнє. Водночас їх точність (90,0 %) і F1-міра (94,1–94,2 %) найнижчі серед усіх моделей, що ілюструє, наскільки ці метрики спотворюються дисбалансом класів.

Ансамблеві методи на основі дерев — випадковий ліс і градієнтний бустинг — досягають майже ідеальної повноти (99,6 % і 99,1 %), але мають найнижчу влучність (91,6 % і 93,3 %) і найнижчу збалансовану точність (63,3 % і 71,1 %). Така поведінка свідчить про схильність цих моделей відносити до синтезованого класу й значну частину справжніх записів, тобто про велику частку хибних відхилень справжнього мовлення за обраного порога. Можливою причиною є відсутність зважування класів для цих моделей.

Класифікатор MLP має високу повноту (98,6 %), але збалансовану точність лише 78,3 % через хибне відхилення понад 40 % справжніх записів. Це свідчить про те, що неглибока нейронна мережа за наявного обсягу навчальних даних не має переваг над простішими моделями на ознаках, сформованих вручну.

Отже, класичні класифікатори помітно відрізняються за балансом між пропуском синтезованого мовлення та хибним відхиленням справжнього. SVM з ядром RBF виявився найкращою класичною базовою моделлю як за збалансованою точністю, так і за F1-мірою.

Ефективність класифікаторів залежно від системи синтезу. У цьому підрозділі ефективність виявлення аналізується окремо для кожної системи синтезу мовлення. На рис. 4 для кожної моделі наведено частку правильно класифікованих тестових записів окремо для справжнього мовлення (real — частка справжніх записів, правильно прийнятих як справжні) і для мовлення, синтезованого ElevenLabs (11labs), RVC, Tortoise та XTTS (частка виявлених синтезованих записів).

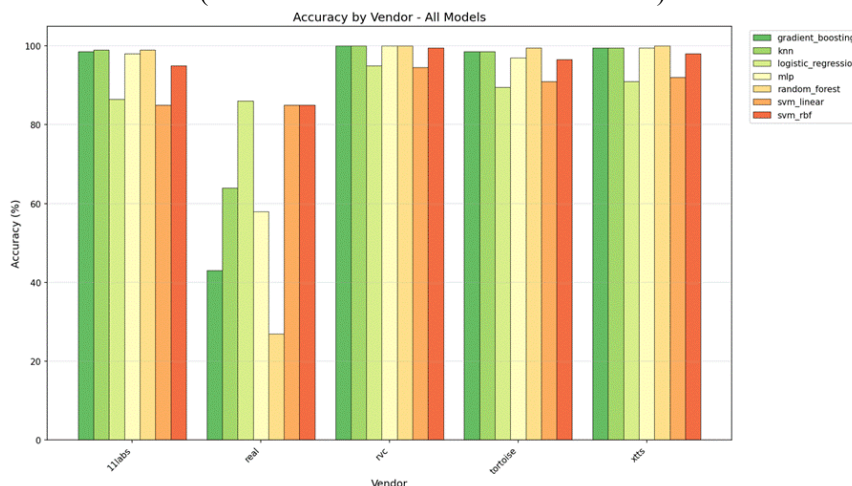


Рис. 4. Точність класифікації окремо для справжнього мовлення та для кожної системи синтезу

Синтезоване мовлення всіх чотирьох систем більшість класифікаторів виявляють з високою точністю. За даними рис. 4, нелінійні моделі (SVM з ядром RBF, випадковий ліс, градієнтний бустинг, KNN і MLP) виявляють 99–100 % записів RVC, 98–100 % записів XTTS, 96–100 % записів Tortoise та 95–99 % записів ElevenLabs. Лінійні моделі мають нижчі значення — від 85 до 95 % залежно від системи синтезу.

Натомість точність на справжньому мовленні значно різниться між класифікаторами і є основним джерелом помилок. Найвищі значення мають логістична регресія (близько 86 %), лінійний SVM і SVM з ядром RBF (близько 85 %). Для KNN і MLP вона становить лише близько 64 % і 58 %, а для градієнтного



бустингу та випадкового лісу — близько 43 % і 27 %, тобто ці моделі хибно відхиляють більшість справжніх записів. Така поведінка узгоджується з високою повнотою та нижчою влучністю цих моделей у табл. 4 і в практичному сценарії розгортання означала б часті відмови легітимним користувачам. Імовірною причиною є дисбаланс класів у навчальній підмножині (1 : 8) у поєднанні з тим, що саме для цих моделей зважування класів не застосовувалося.

Серед систем синтезу найскладнішим для виявлення виявилось мовлення ElevenLabs, однак ця відмінність помітна переважно для лінійних моделей (85–86,5 % проти 89,5–95 % для інших систем) і для SVM з ядром RBF (95 % проти 96,5–99,5 %). Для решти нелінійних моделей точність на ElevenLabs (98–99 %) практично не відрізняється від інших систем. Це може свідчити про те, що мовлення цієї комерційної системи містить менш виражені артефакти в просторі використаних ознак; водночас для узагальнення цього висновку на комерційні системи загалом потрібні дані щонайменше кількох таких систем.

Найзбалансованішу поведінку щодо обох класів демонструє SVM з ядром RBF: близько 85 % правильно прийнятих справжніх записів і не менше 95 % виявлених синтезованих записів для кожної системи синтезу. Лінійні моделі найкраще розпізнають справжнє мовлення, але найбільше залежать від системи синтезу; ансамблеві моделі, KNN і MLP, навпаки, майже не залежать від системи синтезу, проте хибно відхиляють значну частину справжнього мовлення.

Отже, аналіз за системами синтезу показує, що класичні детектори ефективно виявляють мовлення всіх чотирьох досліджених систем, а основним їх обмеженням є хибне відхилення справжнього мовлення. Це підкреслює потребу оцінювати детектори не лише за часткою виявленого синтезованого мовлення, а й за здатністю приймати справжнє, а також використовувати збалансовані метрики й аналізувати результати окремо за класами та системами синтезу.

Порівняння з відомими результатами. Пряме кількісне порівняння з опублікованими результатами ускладнене: у змаганнях ASVspoof основною метрикою є EER, тоді як у цій роботі ефективність оцінено за фіксованого порога, а набори даних і протоколи оцінювання відрізняються. Якісно отримані результати узгоджуються з відомими даними: класичні детектори на ознаках, сформованих вручну, ефективні насамперед тоді, коли атаки відомі на етапі навчання [8], тоді як детектори на основі глибокого навчання та самокерованих представлень демонструють суттєво кращу ефективність і узагальнення [5, 6, 14]. Обчислення EER для досліджених моделей та їх оцінювання на стандартних корпусах є одним із напрямів подальшої роботи.

Обмеження дослідження. Отримані результати слід інтерпретувати з урахуванням таких обмежень. По-перше, записи кожного мовця присутні в усіх підмножинах, а синтезоване мовлення згенеровано для тих самих мовців, тому класифікатори могли частково використовувати характеристики, специфічні для мовців; результати характеризують сценарій із відомими мовцями. По-друге, усі системи синтезу присутні в навчальній підмножині, тобто оцінено виявлення відомих атак; узагальнення на невідомі системи синтезу (зокрема за протоколом «без однієї системи», leave-one-generator-out) не досліджувалося. По-третє, результати отримано на одному розбитті даних без довірчих інтервалів, а справжній клас у тестовій підмножині представлений лише 100 записами, що обмежує точність оцінок частки прийнятих справжніх записів і збалансованої точності. По-четверте, справжні записи (аудіокнижки LibriVox, отримані з платформи YouTube) пройшли стиснення та обробку, характерні для цих джерел, тоді як синтезовані записи такої обробки не зазнавали; тому класифікатори могли частково розрізняти записи за ознаками каналу, а не синтезу. Крім того, у роботі не досліджено стійкість детекторів до шуму, повторного стиснення та передавання каналами зв'язку. Нарешті, інтерпретованість класичних моделей у роботі не досліджувалася і розглядається лише як їх потенційна перевага.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У статті виконано систематичну оцінку класичних методів машинного навчання для виявлення синтезованого мовлення на основі акустичних ознак, сформованих вручну. Для цього сформовано набір даних обсягом 3600 записів, що містить справжнє мовлення 20 мовців і мовлення, синтезоване чотирма сучасними системами синтезу мовлення з тексту та перетворення голосу — RVC, XTTS, ElevenLabs і Tortoise.

Експерименти показали, що за умови, коли мовці та системи синтезу спільні для навчальної і тестової підмножин, класичні класифікатори досягають збалансованої точності від 63,3 до 91,1 %. Найкращою базовою моделлю виявився SVM з ядром RBF (збалансована точність 91,1 %, F1-міра 97,7 %), який забезпечує найзбалансованішу поведінку щодо справжнього та синтезованого мовлення. Лінійні моделі посідають друге й третє місце за збалансованою точністю: вони найкраще розпізнають справжнє мовлення, але частіше пропускають синтезоване. Мовлення, синтезоване системами RVC, XTTS і Tortoise,



нелінійні моделі виявляють майже безпомилково; мовлення комерційної системи ElevenLabs дещо складніше виявити, передусім лінійними моделями.

Водночас виявлено суттєве обмеження класичних детекторів: ансамблеві моделі, KNN і MLP, навчені без зважування класів, хибно відхиляють від 36 до 73 % справжніх записів. Це підкреслює важливість урахування дисбалансу класів, оцінювання з урахуванням порога та ретельного вибору робочої точки під час впровадження засобів протидії підробленню в реальних біометричних системах.

Загалом класичні методи машинного навчання залишаються корисними як обчислювально ефективні базові моделі для виявлення синтезованого мовлення. Проте їх залежність від вибору ознак і потенційна нестійкість до зміни домену зумовлюють потребу в подальших дослідженнях. Подальша робота буде спрямована на обчислення метрик, що не залежать від порога (EER, AUC), оцінювання за протоколами з невідомими мовцями та невідомими системами синтезу, узагальнення між наборами даних і мовами, аналіз важливості ознак, використання самокерованих представлень мовлення та пряме порівняння з наскрізними нейронними детекторами в реалістичних акустичних умовах і каналах передавання.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Li, M., Ahmadiadli, Y., & Zhang, X.-P. (2025). A survey on speech deepfake detection. *ACM Computing Surveys*, 57(7), 1–38. <https://doi.org/10.1145/3714458>
2. Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1), 1–41. <https://doi.org/10.1145/3425780>
3. Sahidullah, M., Kinnunen, T., & Hanilçi, C. (2015). A comparison of features for synthetic speech detection. In *Proceedings of Interspeech 2015* (pp. 2087–2091). ISCA. <https://doi.org/10.21437/Interspeech.2015-472>
4. Lavrentyeva, G., et al. (2017). Audio replay attack detection with deep learning frameworks. In *Proceedings of Interspeech 2017* (pp. 82–86). ISCA. <https://doi.org/10.21437/Interspeech.2017-360>
5. Tak, H., et al. (2021). End-to-end anti-spoofing with RawNet2. In *ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6369–6373). IEEE. <https://doi.org/10.1109/ICASSP39728.2021.9414234>
6. Jung, J.-w., et al. (2022). AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6367–6371). IEEE. <https://doi.org/10.1109/ICASSP43922.2022.9747766>
7. Müller, N., et al. (2022). Does audio deepfake detection generalize? In *Proceedings of Interspeech 2022* (pp. 2783–2787). ISCA. <https://doi.org/10.21437/Interspeech.2022-108>
8. Todisco, M., et al. (2019). ASVspoof 2019: Future horizons in spoofed and fake audio detection. In *Proceedings of Interspeech 2019* (pp. 1008–1012). ISCA. <https://doi.org/10.21437/Interspeech.2019-2249>
9. Yamagishi, J., et al. (2021). ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. In *Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge (ASVspoof 2021)* (pp. 47–54). <https://doi.org/10.21437/ASVspoof.2021-8>
10. Yi, J., et al. (2023). ADD 2023: The second audio deepfake detection challenge. In *Proceedings of the IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023)* (CEUR Workshop Proceedings, Vol. 3597, pp. 125–130). CEUR-WS.org. <https://ceur-ws.org/Vol-3597/>
11. Brydinskiy, V., et al. (2024). Enhancing automatic speech recognition with personalized models: Improving accuracy through individualized fine-tuning. *IEEE Access*, 12, 116649–116656. <https://doi.org/10.1109/ACCESS.2024.3443811>
12. Ravanelli, M., & Bengio, Y. (2018). Speaker recognition from raw waveform with SincNet. In *2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 1021–1028). IEEE. <https://doi.org/10.1109/SLT.2018.8639585>
13. Zaiets, I., et al. (2024). Integrated system for speaker diarization and intruder detection using speaker embeddings. In *Cybersecurity Providing in Information and Telecommunication Systems (CPITS 2024)* (CEUR Workshop Proceedings, Vol. 3654, pp. 228–238). CEUR-WS.org. <https://ceur-ws.org/Vol-3654/>
14. Tak, H., et al. (2022). Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *Proceedings of The Speaker and Language Recognition Workshop (Odyssey 2022)* (pp. 112–119). ISCA. <https://doi.org/10.21437/Odyssey.2022-16>
15. Ruda, K., et al. (2025). Scalability analysis of neural network architectures for voice biometric authentication. In *Cybersecurity Providing in Information and Telecommunication Systems II (CPITS-II 2025)* (CEUR Workshop Proceedings, Vol. 4145, pp. 349–357). CEUR-WS.org. <https://ceur-ws.org/Vol-4145/>



16. Sun, C., Jia, S., Hou, S., & Lyu, S. (2023). AI-synthesized voice detection using neural vocoder artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 904–912).
17. ElevenLabs. (n.d.). *AI voice generator and text to speech*. Retrieved October 9, 2026, from <https://elevenlabs.io>
18. RVC-Project. (2023). *Retrieval-based-Voice-Conversion-WebUI* [Computer software]. GitHub. <https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI>
19. Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning* (Proceedings of Machine Learning Research, Vol. 139, pp. 5530–5540). PMLR.
20. Betker, J. (2023). *Better speech synthesis through scaling* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2305.07243>
21. Casanova, E., et al. (2024). XTTS: A massively multilingual zero-shot text-to-speech model. In *Proceedings of Interspeech 2024* (pp. 4978–4982). ISCA. <https://doi.org/10.21437/Interspeech.2024-2016>

**Ruda Khrystyna**

Doctor of Philosophy, Associate Professor, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0000-0001-8644-411X
khristyna.s.ruda@lpnu.ua

Kos Ihor

Student, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0009-0001-4558-5036
ihor.kos.mkbis.2025@lpnu.ua

Sabodashko Dmytro

Doctor of Philosophy, Senior Lecturer, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0000-0003-1675-0976
dmytro.v.sabodashko@lpnu.ua

Shved Mariia

Candidate of Technical Sciences (PhD), Associate Professor, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0000-0003-0428-7777
mariia.y.shved@lpnu.ua

Hlushchenko Pavlo

Doctor of Philosophy, Assistant, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0000-0002-1262-5484
pavlo.k.hlushchenko@lpnu.ua

DETECTING CLONED VOICES WITH CLASSICAL MACHINE LEARNING

Abstract. The rapid progress of neural text-to-speech (TTS) and voice conversion (VC) technologies has enabled highly realistic synthetic speech, significantly increasing the risk of voice spoofing and social engineering attacks. This work investigates synthetic speech detection using classical machine learning methods based on handcrafted acoustic features. For experimental evaluation, a dataset of 3,600 recordings is constructed: 400 bona fide utterances from 20 speakers and 3,200 utterances synthesized by four modern voice cloning and TTS systems — RVC, XTTS, ElevenLabs, and Tortoise (class ratio 1 : 8). Each recording is represented by a 135-dimensional vector of cepstral, spectral, and temporal features, which is fed to seven classical classifiers, including support vector machines, ensemble methods, and k-nearest neighbors. The best result is obtained with an RBF-kernel SVM, which achieves a balanced accuracy of 91.1%, correctly accepting 85% of bona fide recordings and detecting 97.2% of synthetic ones. Per-system analysis shows that the main source of errors is the false rejection of bona fide speech: for four of the seven models, only 27–64% of bona fide recordings are correctly accepted, whereas synthetic speech is detected with an accuracy of at least 85%. Speech generated by the commercial ElevenLabs system proved the most difficult to detect for linear models. The results constitute baselines for a setting in which speakers and synthesis systems are shared between the training and test sets, and provide a reference point for comparison with deep learning-based detectors.

Keywords: audio deepfake; synthetic speech detection; voice cloning; spoofing countermeasures; classical machine learning; support vector machine; speaker verification.

REFERENCES

1. Li, M., Ahmadiadli, Y., & Zhang, X.-P. (2025). A survey on speech deepfake detection. *ACM Computing Surveys*, 57(7), 1–38. <https://doi.org/10.1145/3714458>



2. Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1), 1–41. <https://doi.org/10.1145/3425780>
3. Sahidullah, M., Kinnunen, T., & Hanilçi, C. (2015). A comparison of features for synthetic speech detection. In *Proceedings of Interspeech 2015* (pp. 2087–2091). ISCA. <https://doi.org/10.21437/Interspeech.2015-472>
4. Lavrentyeva, G., et al. (2017). Audio replay attack detection with deep learning frameworks. In *Proceedings of Interspeech 2017* (pp. 82–86). ISCA. <https://doi.org/10.21437/Interspeech.2017-360>
5. Tak, H., et al. (2021). End-to-end anti-spoofing with RawNet2. In *ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6369–6373). IEEE. <https://doi.org/10.1109/ICASSP39728.2021.9414234>
6. Jung, J.-w., et al. (2022). AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6367–6371). IEEE. <https://doi.org/10.1109/ICASSP43922.2022.9747766>
7. Müller, N., et al. (2022). Does audio deepfake detection generalize? In *Proceedings of Interspeech 2022* (pp. 2783–2787). ISCA. <https://doi.org/10.21437/Interspeech.2022-108>
8. Todisco, M., et al. (2019). ASVspoof 2019: Future horizons in spoofed and fake audio detection. In *Proceedings of Interspeech 2019* (pp. 1008–1012). ISCA. <https://doi.org/10.21437/Interspeech.2019-2249>
9. Yamagishi, J., et al. (2021). ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. In *Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge (ASVspoof 2021)* (pp. 47–54). <https://doi.org/10.21437/ASVSPPOOF.2021-8>
10. Yi, J., et al. (2023). ADD 2023: The second audio deepfake detection challenge. In *Proceedings of the IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023)* (CEUR Workshop Proceedings, Vol. 3597, pp. 125–130). CEUR-WS.org. <https://ceur-ws.org/Vol-3597/>
11. Brydinskiy, V., et al. (2024). Enhancing automatic speech recognition with personalized models: Improving accuracy through individualized fine-tuning. *IEEE Access*, 12, 116649–116656. <https://doi.org/10.1109/ACCESS.2024.3443811>
12. Ravanelli, M., & Bengio, Y. (2018). Speaker recognition from raw waveform with SincNet. In *2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 1021–1028). IEEE. <https://doi.org/10.1109/SLT.2018.8639585>
13. Zaiets, I., et al. (2024). Integrated system for speaker diarization and intruder detection using speaker embeddings. In *Cybersecurity Providing in Information and Telecommunication Systems (CPITS 2024)* (CEUR Workshop Proceedings, Vol. 3654, pp. 228–238). CEUR-WS.org. <https://ceur-ws.org/Vol-3654/>
14. Tak, H., et al. (2022). Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *Proceedings of The Speaker and Language Recognition Workshop (Odyssey 2022)* (pp. 112–119). ISCA. <https://doi.org/10.21437/Odyssey.2022-16>
15. Ruda, K., et al. (2025). Scalability analysis of neural network architectures for voice biometric authentication. In *Cybersecurity Providing in Information and Telecommunication Systems II (CPITS-II 2025)* (CEUR Workshop Proceedings, Vol. 4145, pp. 349–357).
16. Sun, C., Jia, S., Hou, S., & Lyu, S. (2023). AI-synthesized voice detection using neural vocoder artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 904–912).
17. ElevenLabs. (n.d.). *AI voice generator and text to speech*. Retrieved October 9, 2026.
18. RVC-Project. (2023). *Retrieval-based-Voice-Conversion-WebUI* [Computer software]. GitHub. <https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI>
19. Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning* (Proceedings of Machine Learning Research, Vol. 139, pp. 5530–5540). PMLR.
20. Betker, J. (2023). *Better speech synthesis through scaling* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2305.07243>
21. Casanova, E., et al. (2024). XTTS: A massively multilingual zero-shot text-to-speech model. In *Proceedings of Interspeech 2024* (pp. 4978–4982). ISCA. <https://doi.org/10.21437/Interspeech.2024-2016>

Отримано редакцією журналу / Received: 19.06.26

Прорецензовано / Revised: 29.06.26

Схвалено до друку / Accepted: 26.10.26



This work is licensed under Creative Commons Attribution-noncommercial-sharealike 4.0 International License.