Biomedical Engineering and Technology
Issue 12(4), 20223 ISSN (online) 2707-8434

УДК 004.852 + 616.8-005

ПРОГНОЗУВАННЯ РИЗИКУ НАСТАННЯ ІНСУЛЬТУ ЗА ДОПОМОГОЮ ОБРОБКИ НЕЗБАЛАНСОВАНИХ ДАНИХ

Жиляк Максим Євгенович
jilyak88@gmail.com
Городецька Олена Костянтинівна
o.nosovets@gmail.com
кафедра біомедичної кібернетики
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»,
м. Київ, Україна

Реферат *– У контексті медичної науки, інсульт залишається однією з провідних причин смертності та інвалідності, що ставить високі вимоги до ефективності його діагностики та прогнозування. У цьому дослідженні розглядалась проблематика використання незбалансованих даних для прогнозування ризику інсульту, яка є особливо актуальною в умовах гетерогенності симптомів та відсутності універсальних діагностичних методів. Метою даної роботи є вивчення та розробка ефективних прогностичних моделей ризику інсульту, використовуючи сучасні методи машинного навчання, та зосередження на проблемі класового дисбалансу у даних. Основний акцент ставиться на вирішенні викликів, пов'язаних з недостатньою представленістю деяких класів в даних, що є критичним для забезпечення точності прогнозування. Методологія дослідження охоплює декілька етапів: підготовку та обробку даних, використання методів для боротьби з дисбалансом класів (ADAYSN та GAN), а також застосування різних алгоритмів бінарної класифікації. Важливим аспектом є також аналіз впливу різних параметрів на результати прогнозування. Результати дослідження показали, що логістична регресія, навчена на даних, згенерованих за допомогою генеративної нейронної мережі (GAN), продемонструвала найвищу ефективність. Ця модель показала високі показники точності, чутливості, специфічності та зваженої F1-оцінки. Серед аналізованих параметрів особливо значущими виявилися 'is_private_job' (анотація, що пацієнт працює на приватній фірмі), 'is_never_smoked' (анотація, що пацієнт ніколи не курив), та 'is_male' (анотація, що пацієнт чоловічої статі). Загальні висновки дослідження підкреслюють важливість використання методів машинного навчання для прогнозування ризику інсульту, особливо в умовах незбалансованих даних. Вони також вказують на необхідність розробки цілеспрямованих стратегій профілактики, зосереджуючись на ідентифікованих групах ризику, для зниження загальної захворюваності та підвищення ефективності медичних втручань.*
Ключові слова: *інсульт, незбалансовані дані, машинне навчання, ADAYSN, GAN*

I. ВСТУП

Інсульт – це серйозне цереброваскулярне захворювання, яке становить значну загрозу для життя і може призвести до тривалої інвалідності. Ефективність лікування інсульту багато в чому залежить від швидкості діагностики та своєчасного медичного втручання, оскільки зволікання може призвести до незворотних ушкоджень мозку, порушень рухових функцій або навіть смерті.

З огляду на те, що інсульт залишається другою провідною причиною смерті в усьому світі (що підтверджується останніми дослідженнями [1]), актуальність боротьби з цією патологією не викликає сумнівів. В контексті України, до повномасштабного військового вторгнення, статистика свідчила, що смерть від інсульту траплялася приблизно кожні десять хвилин, тобто близько 50 тисяч смертей на рік. Очікується, що з продовженням конфлікту ці цифри значно зростуть. За даними Міністерства охорони здоров'я України, до 2023 року кількість випадків інсульту збільшиться на 16% [2]. Крім того, щороку в Україні реєструється 100-110 тисяч первинних інсультів та 40-50 тисяч повторних інсультів, що призводить до 70-75 тисяч смертей та 20-25 тисяч випадків інвалідності, що робить Україну однією з перших п'яти європейських країн за рівнем захворюваності та смертності від інсульту [3].

Діагностика інсульту є складним завданням, яке вимагає швидкого втручання і ускладняється гетерогенністю симптомів. Традиційні методи діагностики базуються на оцінці клінічних проявів захворювання та аналізі медичних зображень, таких як магнітно-резонансна томографія (МРТ) або

комп'ютерна томографія (КТ) головного мозку. Однак останні технологічні досягнення, особливо в галузі машинного навчання (МН), пропонують перспективні підходи для оптимізації та підвищення точності діагностичних процесів при інсульті [4].

Сучасні дослідження в галузі діагностики інсульту демонструють значний прогрес як з точки зору технологічного розвитку, так і більш глибокого розуміння етіологічних факторів, що сприяють розвитку інсульту. Особливої уваги заслуговує інновація вчених з Університету Південної Каліфорнії [5], які розробили роботизовану систему для оцінки мобільності після інсульту. Ця система використовує роботизовану кінцівку і методологію МН для збору та аналізу тривимірних просторових даних, що дозволяє створити метрику «невикористання руки». Ця інновація надає лікарям унікальну можливість точно оцінити прогрес реабілітації пацієнтів, оскільки система ефективно визначає рівень використання ослабленої кінцівки поза клінічними умовами.

У дослідженні [6] було представлено розробку та верифікацію моделі машинного навчання, призначеної для прогнозування функціонального результату у пацієнтів після гострого ішемічного інсульту. Використовуючи дані багатоцентрового реєстру інсультів, дослідження показало, що найбільш важливими прогностичними факторами є NIHSS (шкала тяжкості інсульту), раннє неврологічне погіршення, вік пацієнта та кількість лейкоцитів.

Дослідження Янкового та ін. [7] було присвячено розробці методології глибокого навчання для оцінки рухової функції у пацієнтів, які перенесли інсульт. У цьому дослідженні використовувалися відеозаписи 23 пацієнтів для класифікації ступеня пов'язаної з інсультом рухової слабкості в лівій руці відповідно до NIHSS. У дослідженні була розроблена бінарна класифікаційна модель, яка показала високу точність (92.1%) у розрізненні помірної та значної рухової слабкості, пов'язаної з інсультом. Крім того, була розроблена трикласова модель, яка продемонструвала точність 89%.

У контексті даної роботи розглядається використання відкритого набору даних пацієнтів [8] для прогнозування ймовірності інсульту. Цей набір даних включає такі параметри, як стать, вік, стан здоров'я, статус куріння та інші важливі фактори. Унікальність цього набору даних полягає в тому, що він використовується в змаганнях з машинного навчання, а частка випадків інсульту в ньому становить близько 5%. Враховуючи відсутність ефективних підходів до подолання цього дисбалансу в даних, було прийнято рішення розробити власний метод прогнозування ризику інсульту, зосередившись, зокрема, на проблемі класового дисбалансу.

II. МЕТА РОБОТИ

Дослідити ризики настання інсульту у пацієнтів, використовуючи методи машинного навчання і обробки незбалансованих даних.

III. ОПИС КЛІНІЧНИХ ДАНИХ

Обраний для дослідження відкритий набір даних [8] охоплює 11 параметрів, які можна класифікувати в чотири основні категорії:

1. Демографічна інформація: ‘gender’ – стать пацієнта (чоловіча, жіноча, або інше); ‘age’ – вік пацієнта (варіюється від 0.08 до 82 років); ‘ever_married’ – статус шлюбу (‘1’, якщо одружений; 0, якщо ні); ‘work_type’ – тип зайнятості (приватна, державна, самозайнятість, пацієнт ніколи не працював, або пацієнт є ще дитиною); ‘residence_type’ – тип резиденції (міська або сільська); ‘smoking_status’ – статус куріння (можливі варіанти, що пацієнт колишній курець, ніколи не курив, курить, або невідомий статус куріння).

2. Анамнез: ‘hypertension’ – чи була у пацієнта гіпертонія (‘0’, якщо ні; ‘1’, якщо так); ‘heart_disease’ – чи було у пацієнта серцево-судинне захворювання (‘0’, якщо ні; ‘1’, якщо так).

3. Клінічна інформація: ‘avg_glucose_level’ – середній рівень глюкози в крові; ‘bmi’ – індекс маси тіла (відношення ваги в кілограмах до квадрату росту в метрах).

4. Таргет (параметр, який необхідно промоделювати): ‘stroke’ – чи наступив у пацієнт через певний момент після збору даних (‘0’, якщо ні; ‘1’, якщо так).

Набір даних містить 5510 пацієнтів, серед яких приблизно 5% зазнали інсульту. Середній вік усієї популяції пацієнтів складає 43.2±22.77 років. Рис. 1 ілюструє віковий розподіл, де відзначено класи таргету ‘stroke’, а також демонструє збільшення частоти випадків інсульту у старших вікових категоріях.

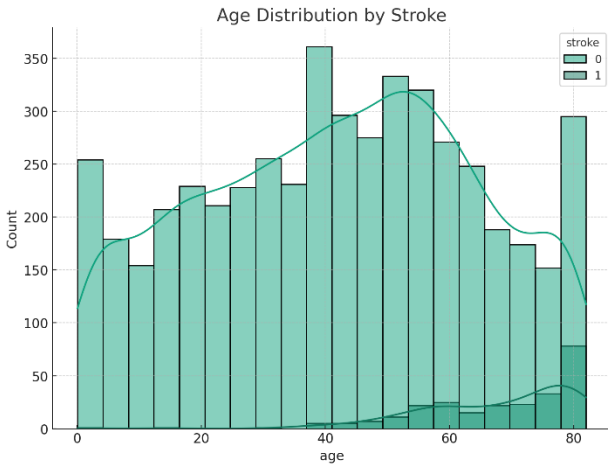

Рис. 1. Вікова гістограма пацієнтів

Зокрема, серед 1628 пацієнтів (29.7% від загальної кількості у наборі даних) віком до 32 років зафіксовано лише 2 випадки інсульту. Крім того, непараметричний критерій Манна-Вітні [9] вказує на статистичну значимість різниці в середньому віці між групами пацієнтів з інсультом та без ($p < 0.05$). Середній рівень глюкози у крові пацієнтів коливається у значному діапазоні, від 55.12 до 271.74 мг/дл, зі середнім значенням 108.85 мг/дл та стандартним відхиленням 44.95 мг/дл. На рис. 2 представлений розподіл середнього рівня глюкози серед пацієнтів, розділений на дві групи.

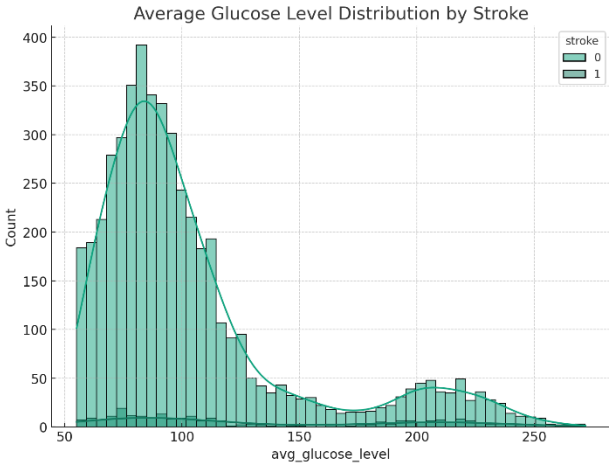

Рис. 2. Гістограма середнього рівня глюкози

Аналіз даних вказує на те, що існує кореляція між підвищенням середнього рівня глюкози та зростанням частоти випадків інсульту. Критерій Манна-Вітні встановив статистичну значимість цієї різниці середніх рівнів глюкози між зазначеними групами ($p < 0.05$), що підкреслює значення рівня глюкози як потенційного фактора ризику інсульту.

Індекс маси тіла (ІМТ) має середнє значення 28.33±6.74 , охоплюючи діапазон від 10.3 до 48. Рис. 3 ілюструє розподіл ІМТ серед пацієнтів, включаючи диференціацію за класом інсульту.

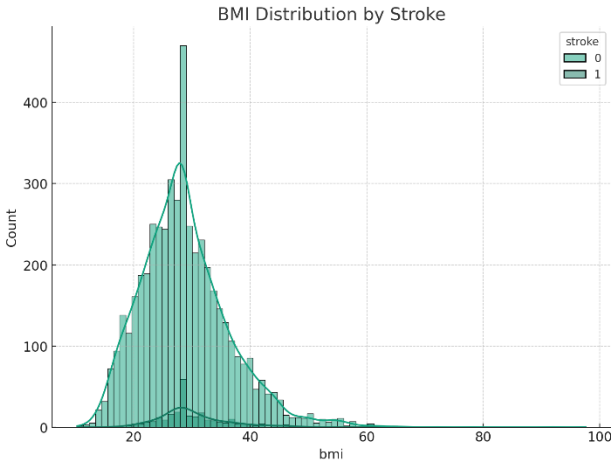

Рис. 3. Гістограма ІМТ

Хоча зв'язок між ІМТ та інсультом виявляється менш виразним, в дослідженні спостерігається тенденція до незначного зростання частоти інсультів у групі осіб з вищим ІМТ. За допомогою непараметричного критерію Манна-Вітні було встановлено статистичну значимість відмінностей у ІМТ

між групами пацієнтів, що перенесли інсульт, та тими, хто не мав інсульту ($p < 0.05$).

У контексті дослідження категоріальних параметрів, отримано низку інсайтів, що відображають зв'язок між певними станами та збільшенням ризику інсульту. Перш за все, було виявлено, що гіпертонія значно підвищує ризик інсульту, з наявністю майже на 9% вищої ймовірності серед осіб, що страждають цим станом. Другий інсайт стосується осіб з серцевими захворюваннями, де ризик інсульту є високим, з виявленням цього стану у 12% пацієнтів з минулою історією серцевих захворювань. Третій аспект пов'язаний із курінням, де спостерігається значний вплив на ризик інсульту, особливо серед колишніх курців. Нарешті, було виявлено, що самозайняті особи мають відносно вищий ризик інсульту, що може бути пов'язано з робочим стресом.

IV. МЕТОДОЛОГІЯ ДОСЛІДЖЕННЯ

У процесі підготовки до розробки прогностичних моделей ризику інсульту, первинна обробка даних відіграє ключову роль. Важливим етапом є перетворення категоріальних параметрів, що не містять числових значень, у формат, придатний для моделювання. Відповідно, бінарні параметри, такі як ‘ever_married’, конвертуються у числові значення ‘0’ або ‘1’. Текстові параметри обробляються за допомогою методу “one-hot encoding” [10], що дозволяє створювати кілька додаткових параметрів. Для ілюстрації, використання “one-hot encoding” для параметру ‘gender’ призводить до формування трьох нових параметрів: ‘is_male’ (значення ‘0’ вказує на відсутність чоловічої статі та ‘1’ - на присутність), ‘is_female’ (аналогічно, ‘0’ для відсутності жіночої статі та ‘1’ - для її присутності) та ‘is_other’ (‘0’ позначає чоловічу або жіночу стать, а ‘1’ - іншу стать).

Наступним кроком у підготовці даних для прогностичного моделювання інсульту є масштабування параметрів. Це стає необхідним через те, що параметри, такі як ‘age’, ‘avg_glucose_level’ та ‘bmi’, мають різні діапазони значень. Для приведення цих параметрів до уніфікованої шкали даних, використовується метод *max-min* нормалізації [11]. Цей метод забезпечує перетворення кожного параметра до діапазону від 0 до 1, що ефективно вирівнює вагу кожного параметра в аналітичному процесі, уникаючи таким чином упередженості щодо окремих параметрів під час моделювання.

Завершальним етапом у процесі підготовки даних для прогностичного моделювання інсульту є розділення даних на тренувальний та тестовий набори. Тренувальний набір використовується для навчання моделей, тоді як тестовий набір призначений для їх об'єктивної оцінки. У зв'язку з обмеженою представленістю випадків інсульту у наборі даних, було обрано співвідношення між тренувальним та тестовим наборами як 9 до 1. Відбір тестової вибірки здійснюється випадковим чином, що допомагає уникнути упередженості моделі до конкретних даних. Для запобігання перенавчання моделі, застосовується 10-fold крос-валідація [12], яка сприяє точному підбору оптимальних гіперпараметрів для прогностичної моделі.

Для моделювання ризику настання інсульту було вибрано п'ять алгоритмів бінарної класифікації. Ці алгоритми включають логістичну регресію [13], яка є відомим інструментом у статистичному аналізі даних, метод опорних векторів (SVM) [14], який широко застосовується для класифікаційних та регресійних завдань, метод екстремального градієнтного підсилювання (XGBoost) [15], що забезпечує ефективне вирішення багатьох задач машинного навчання, метод LightGBM [16], який є ефективним для обробки великих обсягів даних, та метод групового урахування аргументів (МГУА) [17], який застосовується для ідентифікації значущих змінних у багатовимірних даних. Кожен з цих алгоритмів має свої унікальні переваги та особливості, які роблять їх підходящими для різних аспектів моделювання ризику інсульту.

Для вирішення проблеми дисбалансу класів у даних, було обрано два методи:

1. Адаптивна синтетична вибірка (ADAYSN) [18], що ґрунтується на використанні зваженого розподілу для різних

прикладів класів меншості, залежно від їх складності для навчання. Цей метод передбачає генерацію більшої кількості синтетичних даних для випадків класів меншості, які важче навчати, порівняно з тими, що легше піддаються навчанню. ADAYSN сприяє поліпшенню процесу навчання шляхом зменшення зміщення, спричиненого дисбалансом класів, та адаптивним зміщенням межі прийняття рішення про класифікацію на користь складніших випадків.

2. Генеративна нейронна мережа (відома як GAN) для створення синтетичних табличних даних. Архітектура цієї мережі наведена на рис. 4.

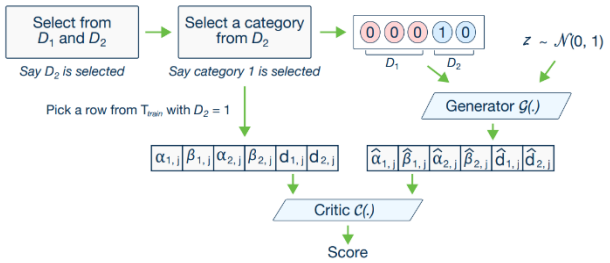

Рис. 4. Архітектура генеративної нейронної мережі

Вона спроектована таким чином, щоб генерувати синтетичні дані, які залежать від одного з дискретних (категоріальних) параметрів. Під час навчання мережі використовується семплування умовних і навчальних даних відповідно до логарифмічної частоти кожної категорії, що забезпечує рівномірне дослідження усіх можливих дискретних значень.

V. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Для оцінки ризику інсульту були застосовані запропоновані п'ять алгоритмів класифікації, кожен з яких був навчений на двох видах синтетичних даних: одні згенеровані з використанням методу ADAYSN, а інші – за допомогою GAN. Результати цих моделей на даних, отриманих за допомогою ADAYSN, представлені в табл. 1. При оцінці ефективності цих моделей були використані класичні метрики, включаючи точність, чутливість та специфічність. Це дозволило глибше оцінити здатність кожного алгоритму розпізнавати випадки інсульту на основі синтетичних даних, забезпечуючи таким чином детальний аналіз їхньої прогностичної потужності.

Таблиця 1 **Результати класифікації на тестовій вибірці (синтетичні дані, згенеровані ADAYSN)**

Метод	Точ.	Чутл.	Спец.
Логістична регресія	74.1%	0.833	0.733
SVM	75.2%	0.806	0.748
XGBoost	78.0%	0.722	0.785
LightGBM	81.2%	0.722	0.819
МГУА	75.8%	0.694	0.763

Результати, отримані від моделей, навчених на синтетичних даних, згенерованих за GAN, представлені в табл. 2.

Таблиця 2 **Результати класифікації на тестовій вибірці (синтетичні дані, згенеровані GAN)**

Метод	Точ.	Чутл.	Спец.
Логістична регресія	66.3%	0.917	0.643
SVM	63.1%	0.917	0.609
XGBoost	90.6%	0.056	0.972
LightGBM	90.2%	0.139	0.961
МГУА	72.9%	0.778	0.725

Ці дані важливі для оцінки ефективності моделювання та забезпечують змогу порівняти вплив різних методів генерації синтетичних даних на якість прогностичних моделей.

Для більш точної оцінки ефективності моделей було обрано використання зваженої F1-оцінки, що враховує дисбаланс класів у наборі даних. Це забезпечує більш справедливий аналіз, оскільки зважена F1-оцінка розглядає як точність, так і повноту моделі, збалансовуючи вплив класів з різною представленістю. Результати цього аналізу представлені у табл. 3.

*Таблиця 3***Порівняння двох методів генерації**

Метод	Зважена F1-оцінка	
	SMOTE	GAN
Логістична регресія	0.823	**0.897**
SVM	0.802	0.895
XGBoost	0.727	0.122
LightGBM	0.729	0.198
МГУА	0.699	0.774

VI. ВИСНОВКИ

В ході дослідження було встановлено, що логістична регресія, навчена на синтетичних даних, згенерованих за допомогою генеративної нейронної мережі (GAN), виявилася найефективнішою у прогнозуванні ризику інсульту. Ця модель показала точність 66.3% на тестовій вибірці, а також високі показники чутливості (0.917) і специфічності (0.643), що свідчить про її здатність правильно ідентифікувати як позитивні, так і негативні випадки. Зважена F1-оцінка моделі склала 0.897, що додатково підтверджує її високу загальну ефективність.

Аналізуючи вплив різних параметрів на модель, було виявлено, що особливо значущими є такі параметри, як ‘is_private_job’, ‘is_never_smoked’ та ‘is_male’. Це вказує на те, що люди, працюючі в приватному секторі, ті, хто ніколи не курили, та чоловіки, мають підвищений ризик інсульту. Ці висновки можуть мати значний вплив на розробку стратегій профілактики інсульту та вказують на необхідність зосередження на цих групах ризику для більш ефективного втручання та зниження загальної захворюваності.

Фінансування. Дане дослідження не отримувало зовнішнього фінансування.

Конфлікт інтересів. Автори заявляють про відсутність конфлікту інтересів.

Згода на публікацію. Усі пацієнти, що мають відношення до рукопису дали згоду на публікацію даної роботи.

ORCID ID та внесок авторів.

1. Makym Zhyliak (70%) – 0009-0006-3730-2442
2. Olena Horodetska (30%) – 0000-0002-8433-3878

ПЕРЕЛІК ПОСИЛАНЬ

[1] Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, Fisher M, Pandian J, Lindsay P. World Stroke Organization (WSO): Global Stroke Fact Sheet 2022. Int J Stroke. 2022 Jan;17(1):18-29. DOI: 10.1177/17474930211065917. Erratum in: Int J Stroke. 2022 Apr;17(4):478. PMID: 34986727.

[2] МОЗ підготувало пропозиції до плану заходів в межах Ukrainian Facility від ЄС [Електронний ресурс] // Сайт Міністерства охорони здоров'я України. – 2023. – Режим доступу до ресурсу: https://bit.ly/47UpHII.

[3] Panteleienko L, Bandrivska S. Neurological letter from Ukraine. Practical Neurology 2023;23:530-535. DOI: 10.1136/pn-2023-003751

[4] Mainali S, Darsie ME and Smetana KS (2021) Machine Learning in Action: Stroke Diagnosis and Outcome Prediction. Front. Neurol. 12:734345. DOI: 10.3389/fneur.2021.734345

[5] Nathaniel Dennler et al. A metric for characterizing the arm nonuse workspace in poststroke individuals using a robot arm.Sci. Robot. 8, eadf7723 (2023). DOI:10.1126/scirobotics.adf772

[6] Lee J, Park KM and Park S (2023) Interpretable machine learning for prediction of clinical outcomes in acute ischemic stroke. Front. Neurol. 14:1234046. DOI: 10.3389/fneur.2023.1234046

[7] I. Yankovyi, S. Yanushkevich, M. Horn and M. Almekhlafi, "Video-based detection of hemiparetic weakness side in post-stroke patient," 2023 IEEE Conference on Artificial Intelligence (CAI), Santa Clara, CA, USA, 2023, pp. 362-363. DOI: 10.1109/CAI54212.2023.00158

[8] Ahmad Hassan, "Stroke Prediction Dataset", IEEE Dataport, doi: 10.21227/mxfb-sc71

[9] McKnight, P.E. and Najab, J. (2010). Mann-Whitney U Test. In The Corsini Encyclopedia of Psychology (eds I.B. Weiner and W.E. Craighead). DOI: 10.1002/9780470479216.corpsy0524

[10] Hancock, J.T., Khoshgoftaar, T.M. Survey on categorical data for neural networks. J Big Data 7, 28 (2020). DOI: 10.1186/s40537-020-00305-w

[11] AKSU, G., GÜZELLER, C. O., & ESER, M. T. (2019). The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model. International Journal of Assessment Tools in Education, 6(2), 170-192. DOI: 10.21449/ijate.479404

[12] R. Malhotra and S. Meena, "Empirical Validation of cross-version and 10-fold cross-validation for Defect Prediction," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2021, pp. 431-438. DOI: 10.1109/ICESC51422.2021.9533030

[13] Alzen, J.L., Langdon, L.S. & Otero, V.K. A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses. IJ STEM Ed 5, 56 (2018). DOI: 10.1186/s40594-018-0152-1

[14] Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, Asdrubal Lopez, A comprehensive survey on support vector machine classification: Applications, challenges and trends, Neurocomputing, Volume 408, 2020, Pages 189-215, ISSN 0925-2312, DOI: 10.1016/j.neucom.2019.10.118

[15] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. DOI: 10.1145/2939672.2939785

[16] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: a highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 3149–3157. DOI: 10.5555/3294996.3295074

[17] Alvin K. Mulashani, Chuanbo Shen, Baraka M. Nkurlu, Christopher N. Mkono, Martin Kawamala, Enhanced group method of data handling (GMDH) for permeability prediction based on the modified Levenberg Marquardt technique from well log data, Energy, Volume 239, Part A, 2022, 121915, ISSN 0360-5442. DOI: 10.1016/j.energy.2021.121915

[18] Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks

(IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 1322-1328. DOI: 10.1109/IJCNN.2008.4633969
[19] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional GAN. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 659, 7335–7345. DOI: 10.5555/3454287.3454946

REFERENCES

[1] Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, Fisher M, Pandian J, Lindsay P. World Stroke Organization (WSO): Global Stroke Fact Sheet 2022. Int J Stroke. 2022 Jan;17(1):18-29. DOI: 10.1177/17474930211065917. Erratum in: Int J Stroke. 2022 Apr;17(4):478. PMID: 34986727.
[2] The Ministry of Health has prepared proposals for an action plan under the EU's Ukrainian Facility [Electronic resource] // Sayt Ministerstva ohorony zdorovya Ukrainy. – 2023. – Access mode to the resource: https://bit.ly/47UpHII
[3] Panteleienko L, Bandrivska S. Neurological letter from Ukraine. Practical Neurology 2023;23:530-535. DOI: 10.1136/pn-2023-003751
[4] Mainali S, Darsie ME and Smetana KS (2021) Machine Learning in Action: Stroke Diagnosis and Outcome Prediction. Front. Neurol. 12:734345. DOI: 10.3389/fneur.2021.734345
[5] Nathaniel Dennler et al. A metric for characterizing the arm nonuse workspace in poststroke individuals using a robot arm.Sci. Robot. 8, eadf7723 (2023). DOI:10.1126/scirobotics.adf772
[6] Lee J, Park KM and Park S (2023) Interpretable machine learning for prediction of clinical outcomes in acute ischemic stroke. Front. Neurol. 14:1234046. DOI: 10.3389/fneur.2023.1234046
[7] I. Yankovyi, S. Yanushkevich, M. Horn and M. Almekhlafi, "Video-based detection of hemiparetic weakness side in post-stroke patient," 2023 IEEE Conference on Artificial Intelligence (CAI), Santa Clara, CA, USA, 2023, pp. 362-363. DOI: 10.1109/CAI54212.2023.00158
[8] Ahmad Hassan, "Stroke Prediction Dataset", IEEE Dataport, doi: 10.21227/mxfb-sc71
[9] McKnight, P.E. and Najab, J. (2010). Mann-Whitney U Test. In The Corsini Encyclopedia of Psychology (eds I.B. Weiner and W.E. Craighead). DOI: 10.1002/9780470479216.corpsy0524
[10] Hancock, J.T., Khoshgoftaar, T.M. Survey on categorical data for neural networks. J Big Data 7, 28 (2020). DOI: 10.1186/s40537-020-00305-w
[11] AKSU, G., GÜZELLER, C. O., & ESER, M. T. (2019). The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model. International Journal of Assessment Tools in Education, 6(2), 170-192. DOI: 10.21449/ijate.479404
[12] R. Malhotra and S. Meena, "Empirical Validation of cross-version and 10-fold cross-validation for Defect Prediction," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2021, pp. 431-438. DOI: 10.1109/ICESC51422.2021.9533030
[13] Alzen, J.L., Langdon, L.S. & Otero, V.K. A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses. IJ STEM Ed 5, 56 (2018). DOI: 10.1186/s40594-018-0152-1
[14] Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, Asdrubal Lopez, A comprehensive survey on support vector machine classification: Applications, challenges and trends, Neurocomputing, Volume 408, 2020, Pages 189-215, ISSN 0925-2312, DOI: 10.1016/j.neucom.2019.10.118
[15] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. DOI: 10.1145/2939672.2939785
[16] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: a highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 3149–3157. DOI: 10.5555/3294996.3295074
[17] Alvin K. Mulashani, Chuanbo Shen, Baraka M. Nkurlu, Christopher N. Mkono, Martin Kawamala, Enhanced group method of data handling (GMDH) for permeability prediction based on the modified Levenberg Marquardt technique from well log data, Energy, Volume 239, Part A, 2022, 121915, ISSN 0360-5442. DOI: 10.1016/j.energy.2021.121915
[18] Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 1322-1328. DOI: 10.1109/IJCNN.2008.4633969
[19] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional GAN. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 659, 7335–7345. DOI: 10.5555/3454287.3454946

Biomedical Engineering and Technology
Issue 12(4), 20223 ISSN (online) 2707-8434

UDC 004.852 + 616.8-005

PREDICTING STROKE RISK VIA HANDLING THE IMBALANCED DATA

Maksym Zhyliak
jilyak88@gmail.com
Olena Horodetska
o.nosovets@gmail.com
Department of Biomedical Cybernetics
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute",
Kyiv, Ukraine

Abstract – *In the context of medical science, stroke remains one of the leading causes of mortality and disability, which places high demands on the effectiveness of its diagnosis and prognosis. This study examined the problem of using unbalanced data to predict stroke risk, which is especially relevant in the context of heterogeneity of symptoms and lack of universal diagnostic methods. The aim of this paper is to study and develop effective predictive models of stroke risk using modern machine learning methods and focus on the problem of class imbalance in data. The main emphasis is placed on solving the challenges associated with the underrepresentation of some classes in the data, which is critical to ensure the accuracy of the prediction. The research methodology covers several stages: data preparation and processing, use of methods to deal with class imbalance (ADAYSN and GAN), and application of various binary classification algorithms. Another important aspect is the analysis of the impact of various parameters on the forecasting results. The results of the study showed that logistic regression trained on data generated by a generative neural network (GAN) demonstrated the highest efficiency. This model demonstrated high accuracy, sensitivity, specificity, and weighted F1 score. Among the analyzed parameters, 'is_private_job' (annotation that the patient works for a private company), 'is_never_smoked' (annotation that the patient has never smoked), and 'is_male' (annotation that the patient is male) were particularly significant. The overall findings of the study emphasize the importance of using machine learning methods to predict stroke risk, especially in the face of unbalanced data. They also point to the need to develop targeted prevention strategies, focusing on identified risk groups, to reduce overall morbidity and increase the effectiveness of medical interventions.*
Keywords *– stroke, imbalanced data, machine learning, ADAYSN, GAN*