УДК 004.8 **DOI:** https://doi.org/10.30837/bi.2026.1(104).08

Н. Б. Гулієв[1], О. С. Назаров[2]
[1]ХНУРЕ, м. Харків, Україна, nural.huliiev@nure.ua, ORCID ID: 0000-0003-2123-0377
[2]ХНУРЕ, м. Харків, Україна, oleksii.nazarov1@nure.ua, ORCID ID: 0000-0001-8682-5000

ДОСЛІДЖЕННЯ СПОСОБІВ УСУНЕННЯ ДИСБАЛАНСУ КЛАСІВ СЕРЕД МЕДИЧНИХ ТА ПСИХОЛОГІЧНИХ ДАНИХ ДЛЯ ПОБУДОВИ АЛГОРИТМУ ВИПАДКОВИЙ ЛІС

Випадковий ліс є широковідомим методом прогнозування, який застосовується в екології, бізнесі, фінансах, медицині, ІТ. Не дивлячись на те, що він є досить потужним механізмом будування відповідних моделей, алгоритм може надавати некоректні результати, тому потребує удосконалення. Таких способів наразі багато, але необхідними в нашому випадку, а саме у спостереженні попереджень розвитку психологічних розладів серед людей, хворих на гіпо- та гіпертиреоз, є не всі. В минулому дослідженні ми обирали найоптимальніший спосіб налаштувань гіперпараметрів для будування дерев рішень, де найпідходящим та найкращим було виявлено метод байєсівської оптимізації. В цьому експерименті наступним кроком є вибір підходу усунення дисбалансу класів серед медичних даних пацієнтів серед таких, як undersampling, oversampling, SMOTE, RUSBoost, збалансований випадковий ліс (BRF) та ADASYN. На основі даних за критеріями часу, accuracy, precision, recall та f1 побудовано лінійну адитивну згортку задля прийняття рішення. За її показниками видно, що в нашому випадку слід обирати RUSBoost в якості способу боротьби з класами меншості, щоб алгоритм видавав більш точні результати.

ВИПАДКОВИЙ ЛІС, ДИСБАЛАНС КЛАСІВ, КЛАС МЕНШОСТІ, ADABOOST, ADASYN, OVERSAMPLING, RUSBOOST, SMOTE, UNDERSAMPLING

N. B. Huliiev, O. S. Nazarov. Research on ways to eliminate class imbalance among medical and psychological data in building a random forest algorithm. Random forest is a widely known forecasting method used in ecology, business, finance, medicine, and IT. Despite being a powerful mechanism for building relevant models, the algorithm can produce incorrect results and therefore needs improvement. There are many such methods available, but not all of them are necessary in our case, namely in monitoring the development of psychological disorders among people with hypo- and hyperthyroidism. In a previous study, we selected the most optimal method of hyperparameter settings for building decision trees, where the Bayesian optimization method was found to be the most suitable and best. In this experiment, the next step is to choose an approach to eliminate class imbalance among patient medical data, such as undersampling, oversampling, SMOTE, RUSBoost, balanced random forest (BRF), and ADASYN. Based on the data, a linear additive convolution was constructed for decision making according to the criteria of time, accuracy, precision, recall, and f1. Its indicators show that in our case, RUSBoost should be chosen as a way to combat minority classes so that the algorithm produces more accurate results.

ADABOOST, ADASYN, CLASS IMBALANCE, MINORITY CLASS, OVERSAMPLING, RANDOM FOREST, RUSBOOST, SMOTE, UNDERSAMPLING

Вступ

Медична галузь стикається з великою кількістю ускладнень, що виникають на тлі вже наявних захворювань. Це стосується й гіпо- та гіпертиреозу, оскільки ці стани можуть провокувати розвиток супутніх патологій, які ускладнюють лікування основної хвороби. Одним із таких ускладнювальних чинників є психологічні порушення.

У реальних умовах набори даних рідко бувають повністю збалансованими, і проблема дисбалансу дедалі частіше ускладнює процес класифікації, що робить її актуальною темою досліджень. Незбалансований датасет означає, що один клас представлений значно більшою кількістю прикладів, тоді як інші — суттєво меншою. Більшість алгоритмів машинного навчання розрахована на приблизно рівномірний розподіл класів. Тому за наявності дисбалансу моделі схильні «віддавати перевагу» більш чисельному класу, демонструючи високу точність саме для нього, тоді як менш представлений клас часто визначається гірше або навіть ігнорується. У результаті це негативно впливає на загальну якість класифікації.

Медичні дані зазвичай формуються на основі інформації про пацієнтів, що нерідко й призводить до дисбалансу. Зокрема, кількість здорових людей у вибірці часто перевищує кількість пацієнтів із певним захворюванням. Крім того, різна поширеність окремих хвороб також зумовлює нерівномірний розподіл класів. Попри те що менші класи зазвичай мають нижчу точність прогнозування, у медицині саме вони становлять найбільший інтерес. Тому помилки при класифікації менш чисельного класу можуть мати значно серйозніші наслідки, ніж помилки щодо більшості [1].

Об’єкт даного дослідження є саме процес прогнозування психологічних розладів у пацієнтів із гіпо- та гіпертиреозом на основі медичних і психологічних показників.

Предмет дослідження – методи усунення дисбалансу класів даних алгоритму Random Forest.

1. Дослідження оптимізації алгоритму випадковий ліс для аналізу даних пацієнтів

Random Forest – це метод ансамблевого навчання, який застосовує дерева рішень задля того, щоб згенерувати більш точні та надійні прогностичні результати. Кожне з дерев будується за рахунок випадкових підмножин ознак вибірки та бутстреп-множини навчальних даних. Остаточний результат визначається після процесу голосування серед них. Перевагою алгоритму є можливість зменшення ступеня перенавчання та ефективність у дослідженнях з великим обсягом вхідних атрибутів вибірки.

Алгоритм випадкового лісу демонструє достатню стійкість до шуму та аномальних значень, адже кожне дерево будується на різних підвибірках об'єктів і ознак. Завдяки цьому окремі нетипові або екстремальні спостереження мають менший вплив на підсумковий результат моделі.

Дерева рішень є одним із найпоширеніших підходів до побудови алгоритмів класифікації в сучасних інформаційних системах завдяки їх численним перевагам. Зокрема, вони дозволяють працювати як з числовими, так і з категоріальними ознаками, демонструють високу стійкість до аномальних або екстремальних значень спостережень, а також забезпечують наочну інтерпретацію отриманих правил прийняття рішень. Це робить їх корисними для пояснення закономірностей у даних та застосування в експертних системах. Однією з ключових задач, що вирішуються за допомогою експертних систем, є класифікація — процес визначення класу об'єкта шляхом зіставлення його з одним із відомих системі класів.

Попри усе вищезгадане, даний метод алгоритму має низку недоліків. Одним із таких є дисбаланс класів, що і є темою даного дослідження.

У практичних задачах дисбаланс класів є досить поширеним явищем, особливо в тих сферах, де один клас суттєво переважає інші за кількістю прикладів. У таких випадках моделі машинного навчання зазвичай «схиляються» до класу більшості, що погіршує їхню здатність коректно розпізнавати об'єкти менш представленого класу. Це особливо небезпечно в галузях із високим рівнем ризику — зокрема в медицині, системах виявлення фінансового шахрайства чи прогнозуванні надзвичайних подій. Помилки у визначенні рідкісних захворювань, шахрайських операцій або екстремальних погодних явищ можуть мати серйозні наслідки. Наприклад, невчасне виявлення рідкісної хвороби може затримати лікування пацієнта, а пропущені випадки шахрайства — спричинити значні фінансові збитки. Тому подолання дисбалансу класів є не лише технічною задачею, а й важливою умовою забезпечення справедливості, надійності та точності систем ухвалення рішень.

Причиною виникнення дисбалансу класів є велика перевага кількості одного з класів поміж інших, тому такий клас вважається негативним. Взагалі, передбачається, що в задачах класифікації розподіл даних збалансований. Тому в випадку оберненої ситуації клас із меншою кількістю вибірки ігнорується під час аналізу або досліджується некоректно [2].

2. Матеріали і методи досліджень

Розв'язком багатокритеріальних задач є пошук кращого варіанту серед можливих на основі поставлених вимог. В таких випадках застосовують два види множини методів. Перший зменшує кількість критеріїв, а друга – кількість альтернатив. Для цього експерименту краще використати перший, який має такі різновиди: метод відстані, згортки, головні критерії та граничні точки.

Методи згорток поділяються на лінійні, адитивні, мультиплікативні та максимінні. Метою застосування згорток є узагальнення усіх критеріїв аналізу.

Адитивна розраховується за наступною формулою:

$$K(x)=\sum_{j=1}^{n} a_j K_j(x), \tag{1}$$

де $K(x)$ – загальний критерій для альтернативи $x \in X$, $(K\_1(x), ..., K\_j(x), ..., K\_n(x))$ – набір вихідних критеріїв; n – число вихідних критеріїв; $a\_j(x)$ – нормуючий множник, який вказує на вагу альтернативи.

Найкращий із усіх можливих альтернатив задачі обчислюється за допомогою наступної формули:

$$x^n = \arg\max_{x \in X} K(x), \tag{2}$$

Тобто результатом є найбільше значення, отримане методом згортки.

Мультиплікативна згортка розраховується за допомогою такої формули:

$$K(x)=\prod_{j=1}^{n} K_j^{a_j}(x). \tag{3}$$

Максимінна згортка знаходиться за формулою:

$$K(x)=\max_i \min_j a_{ij} K_j(x). \tag{4}$$

Найкращі результати за мультиплікативними та максимінними згортками обчислюються за формулою (2).

Метод граничних критеріїв застосовується в задачах проектування і планування, в яких порогові значення критеріїв набувають значень $k\_j(x) \geq k\_j 0; j = 1 ..., n$. Формула обчислення цього способу наступна:

$$K(x) = \min_j \left(\frac{K_j(x)}{K_{jo}(x)}\right) \tag{5}$$

Найкращий результат обирається формулою 2.

Метод відстані використовує відстань, яка є додатковою метрикою. Наприклад, для вибору ідеального рішення цілком достатньо інформації. Обчислимо відстань до значення максимуму $d(x)$ для кожної альтернативи. Тоді найкраща альтернатива буде відомою із застосуванням формули:

$$x^* = \arg\min_{x \in X} d(x). \tag{6}$$

Навіть обравши метод із першої множини, доцільним може бути застосування принципу Парето (різновид другої вибірки), який полегшує подальше дослідження виключенням наявних альтернатив, якщо усі їхні значення критеріїв менші за значення інших варіантів.

Може бути інший випадок, коли значення неконтрольовані, що ускладнює побудову моделі, тому при цьому використовують метод гарантованого результату.

Для даного дослідження краще обрати метод згортки, бо порогові показники складно вирахувати. Найпростішою та точно підходящою буде лінійна адитивна згортка, якщо буде необхідним – також принцип Парето.

Першим кроком є вибір критеріїв, за якими будуть порівнюватися варіанти.

Другий крок умовний: дані можуть бути кількісними, якими лінійна адитивна згортка коректно оперує, але якщо вони – якісні, то необхідно конвертувати їх у кількісні, що змінить першочергову таблицю даних дослідження.

Якщо значення знаходяться в різних мірах вимірювання або проміжках, необхідно нормалізувати дані.

Під час четвертого етапу деякі з альтернатив видаляються методом Парето: цей принцип передбачає виключення варіантів, якщо вони прозоро програють за усіма своїми показниками критеріїв з-поміж інших альтернатив.

П'ятий крок – ранжування значень, що означає розрахунок вагових коефіцієнтів кожного з критеріїв.

Шостим та останнім кроком є обчислення значення згортки для кожної із альтернатив: знаходимо суму добутків пар значень критеріїв та їх вагових коефіцієнтів.

Задачею дослідження є прийняття рішення з приводу вибору найкращого способу видалення дисбалансу даних при реалізації методу випадковий ліс.

3. Аналіз літературних джерел

В одному дослідженні задачею було обрати найоптимальніший метод ансамблевого навчання для видалення або принаймні зменшення дисбалансу класів даних за допомогою таких методів, як збалансований випадковий ліс (BRF), SMOTE-випадковий ліс або SMOTE-RF, RUSBoost та SMOTEBoost. Експеримент брав до уваги також AdaBoost та традиційний алгоритм випадкового лісу. Було застосовано 13 наборів множин з різними ступенями дисбалансу класів, кожна з яких мала бінарні класифікаційні значення. Тестова вибірка склала 20%, а навчальна – 80%. Для підтримки пропорційності класів використовували стратифіковану множину. Кожному із вище зазначених способів передувала оптимізація шляхом різного налаштування гіперпараметрів за 10 разів. Найкращий варіант обирався, опираючись на значення часу обчислення, точності та відтворення. Метод збалансованого випадкового лісу показав найвищі показники відтворення та точності, а звичайний алгоритм виграв у значенні ефективності моделі [3].

У будь-якому бізнесі необхідно проводити аналіз існуючих клієнтів задля розрахунку прогнозування спадання їхньої зацікавленості у продукції. Існує дослідження, в якому розглядалося дане питання стосовно банківської галузі, де також був дисбаланс класів даних, що вирішувалося за допомогою таких способів, як RUS – метод випадкової недостатньої вибірки, ROS – метод випадкової надмірної вибірки та SMOTE – метод синтетичної надмірної вибірки меншин. Дані включали 10 000 записів з 14 характеристиками та були оброблені алгоритмами LASSO Logistic Regression та Random Forest. Ефективність моделей досліджувалася на основі таких показників, як f1, точності, відтворюваності та прецизійності. Random Forest у поєднанні із ROS-методом показав найвищий результат 86% точності, із SMOTE – 82% та із RUS – 79%. А логістична регресія відповідно 71%, 71% та 73%. В даному випадку кращим виявився спосіб випадкової надмірної вибірки [4].

Алгоритм випадкового лісу є одним із широко використовуваних методів класифікації машинного навчання, оскільки має перевагу у вигляді зменшення ризику перенавчання та покращення загальної ефективності прогнозування. Однак для даних із незбалансованими класами цей алгоритм не дозволяє досягти найкращої ефективності, особливо у прогнозуванні даних у класі меншості. Як результат, у цій статті пропонуються два підходи до повторної вибірки для збалансування даних: техніка синтетичної надмірної вибірки меншості (SMOTE) та техніка синтетичної надмірної вибірки меншості з редагованими найближчими сусідами (SMOTE-ENN). Для техніки класифікації даних алгоритм випадкового лісу застосовується до вихідних даних, а потім до результатів повторної вибірки з використанням як SMOTE, так і SMOTE-ENN. Приклад було застосовано до даних про затримку росту, що складалися з 421 випадку в класі більшості та 79 у класі меншості. Було отримано точність 89% для вихідних даних, 90% для даних, передискретизованих за допомогою SMOTE-ENN, та 91% для даних, передискретизованих за допомогою SMOTE. Найкраща точність була отримана за допомогою техніки передискретизації SMOTE, однак вона не була особливо значущою [5].

Oversampling вирішує проблему дисбалансу класів шляхом генерації даних у класі меншості для покращення ефективності класифікації. Він був застосований, щоб не допустити втрати даних у незбалансованій вибірці. Методи надмірної вибірки меншого з класів SMOTE та ADASYN генерують дані та застосовують принцип суміжності, а також зменшують вірогідність виникнення перенавчання. У даному дослідженні

Random Forest поєднали з цими способами боротьби з дисбалансом класів задля аналізу даних (9 атрибутів та 1 бінарний залежний показник) економічної кризи в Індонезії. Target-значення експерименту представляє собою значення того, чи наявні кризові умови, чи ні. Результати свідчили про те, що обробка дисбалансу класів дійсно приводила до кращих результатів, а саме за допомогою методу ADASYN з такими значеннями Accuracy, Recall, Precision, F1 та ROC AUC, як 98.26%, 66.67%, 72.22%, 65.57% та 82.93% відповідно [6].

Як відомо, однією із головних причин смерті є захворювання серця, тому дослідження в цьому питанні тривають постійно. Особливо результативними виявилися моделі LASSO Logistic Regression, Support Vector Machine та Random Forest у спостереженнях аналізу ризиків та профілактичних процедур. Під час останніх досліджень виявлено, що побудова моделей була незначно оптимізована за допомогою SMOTE, Random Oversampling та Random Undersampling. Проводився новий експеримент, метод якого було відстежити ефективність остаточних методів прогнозування серцево-судинних захворювань із застосуванням вищезгаданих методів усунення дисбалансу класів. В результаті встановлено, що кращим варіантом була розробка моделі Random Forest із SMOTE, що було видно в показниках точності, специфічності та чутливості [7].

Автоматична ідентифікація структури мозку за допомогою магнітно-резонансної томографії є дуже важливою як для досліджень у галузі нейробіології, так і як можливий інструмент клінічної діагностики. У цьому дослідженні представлено нову стратегію повністю автоматизованої сегментації гіпокампу за допомогою МРТ. Вона базується на алгоритмі з контролем, який називається RUSBoost, і поєднує випадкову підбірку даних з алгоритмом підсилення. RUSBoost – це алгоритм, спеціально розроблений для незбалансованої класифікації, який підходить для великих наборів даних, оскільки використовує випадкову підбірку більшості класів. Ефективність RUSBoost порівнювали з ефективністю ADABoost, Random Forest та загальнодоступного пакета сегментації мозку FreeSurfer. Це дослідження було проведено на наборі даних із 50 структурних зображень мозку з вагою T1. Незалежний набір даних із 50 структурних сканів мозку з вагою T1 був використаний для незалежної валідації повністю навчених стратегій. Знову сегментації RUSBoost вигідно відрізнялися від ручних сегментацій, маючи найвищі показники серед чотирьох інструментів. Більше того, коефіцієнт кореляції Пірсона між об'ємами гіпокампу, обчисленими за допомогою ручної сегментації та сегментації RUSBoost, становив 0,83 (0,82) для лівої (правої) сторони, що є статистично значущим і вищим за показники, обчислені за допомогою Adaboost, Random Forest та FreeSurfer. Запропонований метод може бути придатним для точної, надійної та статистично значущої сегментації гіпокампу [8].

Розглянуто чимало існуючих методів усунення дисбалансу класів, але варто визначити, який саме підходить поставленій задачі.

4. Експериментальні дослідження

Проведемо дослідження та оберемо найпідходящий спосіб усунення дисбалансу класів для алгоритму випадковий ліс, написаного задля аналізу медичних та психологічних показників.

У дослідженні альтернативами виступатимуть наступні:

- random undersampling;
- random oversampling;
- SMOTE;
- RUSBoost (Random Undersampling + AdaBoost);
- збалансований випадковий ліс (BRF);
- ADASYN.

Критеріями розгляду, за якими будуть будуватись три моделі, будуть такі атрибути, як:

- age – вік,
- sex – стать,
- on_thyroxine – чи приймає тироксин,
- query_on_thyroxine – запит на тироксин,
- on_antithyroid_meds – чи приймає антитиреоїдні ліки,
- sick – чи хворий,
- pregnant – вагітність,
- thyroid_surgery – чи робилась операція на щитоподібній,
- I131_treatment – лікування радіоактивним йодом,
- query_hypothyroid – підозра на гіпотиреоз,
- query_hyperthyroid – підозра на гіпертиреоз,
- lithium – чи приймає літій,
- goitre – зоб,
- tumor – пухлина,
- hypopituitary – гіпопітуїтаризм,
- psych – чи є психічні розлади,
- TSH_measured – чи вимірювався TSH,
- TSH – значення TSH,
- T3_measured – чи вимірювався T3,
- T3 – значення T3,
- TT4_measured – чи вимірювався TT4,
- TT4 – значення TT4,
- T4U_measured – чи вимірювався T4U,
- T4U – значення T4U,
- FTI_measured – чи вимірювався FTI,
- FTI – значення FTI,
- TBG_measured – чи вимірювався TBG,
- TBG – значення TBG,
- referral_source – джерело направлення,
- target – цільовий клас,
- patient_id – ID пацієнта.

Написаний код на Python показав, що шість алгоритмів мають такі показники (див. табл. 1).

Таблиця 1

Числові характеристики алгоритмів

Методи та критерії	Accuracy	Precision	Recall	F1	Час
Random under-sampling	0.535754	0.840746	0.535754	0.610383	0,0457
Random over-sampling	0.944794	0.943255	0.944794	0.942957	3,2543
SMOTE	0.941089	0.938677	0.941089	0.939295	13,3661
RUSBoost	0.719155	0.545685	0.719155	0.620273	0,0301
Збалансований випадковий ліс (BRF)	0.696184	0.882207	0.696184	0.750664	0,1383
ADASYN	0.938496	0.935228	0.938496	0.936277	13,1897

Як бачимо, показник часу потребує нормалізації, щоб точно оцінити його, адже чим більше час, тим гірше. Перебудуємо таблицю (див. табл. 2).

Таблиця 2

Змінені числові характеристики алгоритмів

Методи та критерії	Accuracy	Precision	Recall	F1	Час
Random under-sampling	0.535754	0.840746	0.535754	0.610383	21,8818381
Random over-sampling	0.944794	0.943255	0.944794	0.942957	0,30728575
SMOTE	0.941089	0.938677	0.941089	0.939295	0,07481614
RUSBoost	0.719155	0.545685	0.719155	0.620273	33,2225914
Збалансований випадковий ліс (BRF)	0.696184	0.882207	0.696184	0.750664	7,23065799
ADASYN	0.938496	0.935228	0.938496	0.936277	0,07581674

Принцип Парето в даному випадку неможливо застосувати, тому усі альтернативи залишаються.

Тепер розрахуємо значення згортки для кожного із варіантів (див. табл. 3).

Таблиця 3

Значення згортки

Методи	Згортка
Random under-sampling	2,59140864
Random over-sampling	2,43651573
SMOTE	2,41595685
RUSBoost	3,19695142
Збалансований випадковий ліс (BRF)	2,25584928
ADASYN	2,40859136

Найкращий результат належить алгоритму RUSBoost [9-10].

Зображення роботи методу наведено нижче (рис. 1).

RUSBoost — це алгоритм, створений спеціально для задач із незбалансованими класами, який добре підходить для роботи з великими обсягами даних, оскільки базується на випадковій недовибірці об'єктів класу більшості.

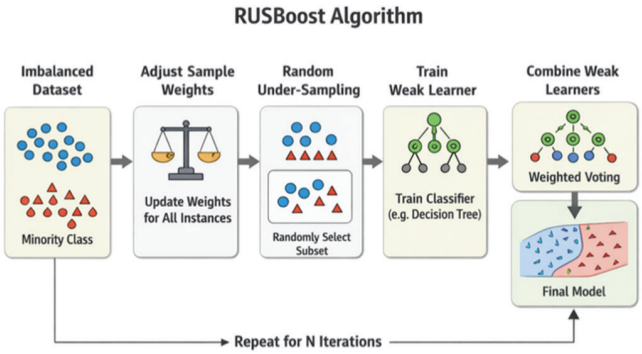

Рис. 1. RUSBoost

Метод поєднує випадкове зменшення кількості прикладів більш представленого класу з механізмом boosting. Він є альтернативою SMOTEBoost, що використовує надвибірку, генеруючи нові зразки меншості шляхом інтерполяції між наявними прикладами та поєднуючи цей процес із підсиленням. Через створення синтетичних даних SMOTEBoost збільшує тривалість навчання моделі. Хоча цей підхід продемонстрував ефективність у низці застосувань, переважно для наборів даних невеликого розміру, зі зростанням обсягу навчальних даних обчислювальні витрати SMOTE істотно підвищуються, що може зробити його непрактичним [11].

У випадку дуже великих датасетів, зокрема 3D MRI, вибір стратегії балансування стає критично важливим. Використання недовибірки, як у RUSBoost, дозволяє значно скоротити час навчання порівняно з методами надвибірки. Водночас основним недоліком цього підходу є можлива втрата частини інформації через видалення окремих прикладів із навчальної вибірки.

Крім того, існують результати досліджень, які свідчать, що RUSBoost може перевершувати SMOTEBoost, будучи простішим і швидшим методом та нерідко забезпечуючи кращі показники класифікації. Наскільки відомо авторам, це перше застосування RUSBoost-класифікаторів для задачі сегментації гіпокампа [12].

Висновки

Алгоритм випадковий ліс вирізняється своєю ефективністю, тому відомий та застосовується в багатьох сферах, але все одно потребує оптимізації. В даному дослідженні аналізувалася проблема дисбалансу класів даних.

Досі цікаво проводити дослідження для розв'язання проблеми усунення дисбалансу класів задля удосконалення роботи алгоритмів. Одним із таких популярних дієвих методів залишається RUSBoost-підхід, який і сам потребує покращення, тому триває робота над видаленням або зменшенням його недоліків.

Розглянуто існуючі методи в різних сферах, а саме: undersampling, oversampling, SMOTE, RUSBoost, збалансований випадковий ліс (BRF) та ADASYN. Метою дослідження було вирішити багатокритеріальну задачу вибору найпідходящого способу усунення класу меншості. Тому методом дослідження обрано лінійну адитивну згортку, яка надала найкращий результат на основі таких показників, як час роботи алгоритму, accuracy, precision, recall та f1. В результаті виявлено, що для усунення дисбалансу класів психологічних та медичних даних слід застосовувати RUSBoost-підхід.

RUSBoost як підхід до балансування даних був запропонований Seiffert і колегами для зменшення обчислювального навантаження, притаманного SMOTEBoost. Метод поєднує випадкову недовибірку (Random Under Sampling, RUS) із алгоритмом підсилення AdaBoost.

Для вирівнювання розподілу класів застосовується стратегія RUS, що полягає у випадковому вилученні частини зразків із класу більшості. Кожна ітерація boosting складається з двох кроків. Спочатку виконується недовибірка для всіх класів, за винятком найменш представленого, і вона триває доти, поки кількість прикладів у кожному класі не стане однаковою та не дорівнюватиме розміру класу меншості. Після цього до сформованої збалансованої підвибірки застосовується алгоритм AdaBoost. Отже, на кожній ітерації використовується оновлена навчальна вибірка, а не фіксований набір даних [13-14].

До недоліків RUSBoost належить потенційна втрата інформації через видалення частини прикладів під час недовибірки. Попри це, алгоритм ефективно застосовується як у задачах бінарної, так і багатокласової класифікації.

За результатами лінійної адитивної згортки RUSBoost має найвищий показник, але варто провести повторне дослідження із ваговими коефіцієнтами важливості критеріїв задля більш точного вибору методу боротьби із дисбалансом класів у дереві рішень [15-16].

У цьому дослідженні було вивчено ефективність різних методів класифікації на наборах даних з різним рівнем дисбалансу класів, що дозволило досягти мети дослідження – визначити оптимальні методи обробки дисбалансованих даних.

Проведено системне порівняння методів балансування (SMOTE, ADASYN, RUS, ROS, RUSBoost, balanced RF) для задачі прогнозування психологічних розладів серед пацієнтів із тиреоїдними порушеннями.

Запропоновано підхід до попередження розвитку психологічних розладів шляхом ранньої стратифікації ризику на основі ансамблевих методів машинного навчання з урахуванням дисбалансу класів.

Запропоновано модель машинного навчання для раннього прогнозування психологічних розладів у пацієнтів із гіпо- та гіпертиреозом на основі комплексного аналізу ендокринних та психометричних показників.

Список літератури

[1] Buda, M., Maki, A., & Mazurowski, M.A. (2021). A multiple combined method for rebalancing medical data with class imbalance. Computers in Biology and Medicine, 135, 104589. DOI: https://doi.org/10.1016/j.compbiomed.2021.104589.

[2] Prasetyo, E., et al. (2018). Evaluating Ensemble Learning Techniques for Class Imbalance Problem. Scientific Journal of Informatics, 5(2), pp. 184–193. URL: https://journal.unnes.ac.id/journals/sji/article/view/15937/2440

[3] Implementation of Imbalanced Learning Methods Using RUSBoost. Mortalita, 4(2). URL: https://ejournal.darunnajah.ac.id/index.php/mortalita/article/view/709/390

[4] Comparative Analysis of Imbalanced Data Handling Using Ensemble Methods. RESTIA, 7(3). URL: https://journal.aiskauniversity.ac.id/index.php/restia/article/view/1906/853

[5] Performance Analysis of RUSBoost for Imbalanced Dataset Classification. International Journal of Multidisciplinary and Current Research, 11. URL: https://ijmcr.in/index.php/ijmcr/article/view/1121/855

[6] Application of Ensemble Learning Methods on Imbalanced Medical Dataset. Jurnal Matematika, Statistika dan Komputasi, 20(1). URL: https://journal.unhas.ac.id/index.php/jmsk/article/view/35552/12001

[7] Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., & Herrera, F. (2018). Learning from Imbalanced Data Sets. Journal of Information and Visualization, 8(3). URL: https://joiv.org/index.php/joiv/article/view/2283

[8] Iglesias, J.E., Liu, C.Y., Thompson, P.M., & Tu, Z. (2015). Automated hippocampal segmentation in 3D MRI using random undersampling with boosting algorithm. Pattern Analysis and Applications, 18, pp. 851–864. DOI: https://doi.org/10.1007/s10044-015-0492-0

[9] Kaur, H., Pannu, H.S., & Malhi, A.K. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. Journal of Big Data, 7, Article 65. DOI: https://doi.org/10.1186/s40537-020-00349-y.

[10] U. Hasanah, A. M. Soleh, and K. Sadik, Effect of Random Under Sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models, Jurnal Matematika, Statistika dan Komputasi 21 (2024) 88-102.

[11] R. Hidayat, M. A. Syawaludin, and N. Nurmalitasari, Prediksi Churn Pelanggan Multinational Bank Menggunakan Algoritma Machine Learning, Simpatik: Jurnal Sistem Informasi dan Informatika 4 (2024) 89-97.

[12] F. Ismail and I. I. Lawanda, Implementasi EDMS dalam Penataan Dokumen di Rail Document System PT. Kereta Api Indonesia (Persero) Daerah Operasi 1 Jakarta, Baca: Jurnal Dokumentasi Dan Informasi 41 (2020) 143-168.

[13] S. M. Kim, Y. Kim, K. Jeong, H. Jeong, and J. Kim, Logistic LASSO Regression for the Diagnosis of Breast Cancer Using Clinical Demographic Data and the BI-RADS Lexicon for Ultrasonography, Ultrasonography 37 (2018) 36-42.

[14] M. Marcellina and A. Mukhlason, Analisis Prediktif Churn untuk Meningkatkan Tingkat Retensi Pelanggan pada Perusahaan SaaS Menggunakan Machine Learning, ILKOMNIKA 6 (2024) 21-32.

[15] R. Zhu, Y. Guo, and J.-H. Xue, "Adjusting the imbalance ratio by the dimensionality of imbalanced data," Pattern Recognit. Lett., vol. 133, pp. 217–223, 2020, doi: https://doi.org/10.1016/j.patrec.2020.03.004.

[16] M. Çakır, A. Degirmenci, and O. Karal, "Exploring the Behavioural Factors of Cervical Cancer Using ANOVA and Machine Learning Techniques BT -Science, Engineering Management and Information Technology," A. Mirzazadeh, B. Erdebilli, E. Babaee Tirkolaee, G.-W. Weber, and A. K. Kar, Eds., Cham: Springer Nature Switzerland, 2023, pp. 249–260

Received (Надійшла) 14.01.2026

Accepted for publication (Прийнята до друку) 10.02.2026

Publication date (Дата публікації) 27.03.2026