**УДК 004.622:004.85**

*Кучер М.О., студент 2 курсу магістратури, спеціальність 122 «Комп'ютерні науки»*
*Бабаков Р.М. к.т.н., доцент, доцент кафедри інформаційних технологій*

# МЕТОДИ БОРОТЬБИ З ДИСБАЛАНСОМ КЛАСІВ В ДАТАСЕТАХ ДЛЯ НАВЧАННЯ НЕЙРОННИХ МЕРЕЖ

*Донецький національний університет імені Василя Стуса, м. Вінниця*

На сьогодні все частіше використовуються нейронні мережі в всіх сферах життя людини, проєкти на їх основі дають можливість автоматизувати чи полегшити ті задачі, які вважались недоступними для комп'ютера і були прерогативою виключною людини. Але часто перешкодою в створення якісної моделі нейромережі є поганий вибір даних для навчання, одною з проблем в даних є дисбаланс класів який представляє собою значну різницю представлених класів яка виникає через внутрішні фактори, такі як природній розподіл даних, а саме медичні діагнози, коли більшість пацієнтів здорові, або ж зовнішні, що виникають через процедури збору чи зберігання даних [1].

Оскільки чутливість для дисбалансу класів зростає зі збільшенням складності моделі, для деяких даних уникнення дисбалансу істотно впливає на результат навчання нейронної мережі [2].

Розглянемо набір даних «Ships in Satellite Imagery»[3] та спробуємо проаналізувати наявність дисбалансу, розглянемо методи, які дозволяють боротися з ним.

Наведемо короткий опис даних в даному наборі, датасет представляє собою набір з сцен, що є супутниковими знімками поверхні землі з різними об'єктами, а також набір відсортованих зображень маленького розміру з прикладами кораблів і зображення всього іншого, що не можна вважати кораблем: вода, будівлі, шматки кораблів, уламки.

Всього датасет містить майже тисячу зображень кораблів і три тисячі інших зображень.

Завантаживши зображення і вивівши співвідношення в вигляді діаграми отримали результат, який продемонстровано на рисунку 1.

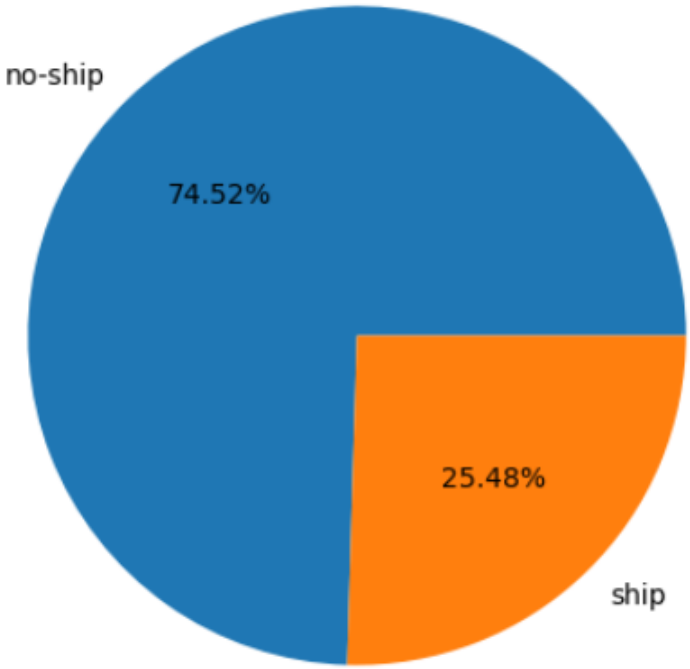


Рисунок 1 – Результат виведення співвідношення класів

Через дисбаланс досліджуваних двох класів навчання моделі може бути викривлене, так як мережа буде частіше схилятись до передбачення в бік відсутності корабля, через те, що звичайним вгадуванням, не дивлячись на малюнок можна досягти приблизно 75% точності.

Даний бисбаланс можна усунути аугментацією даних, тобто дублюванням меншого з двох класів до тих пір, поки обидва класи не будуть приблизно рівні, після чого діаграма буде виглядати так, як продемонстровано на рисунку 2.

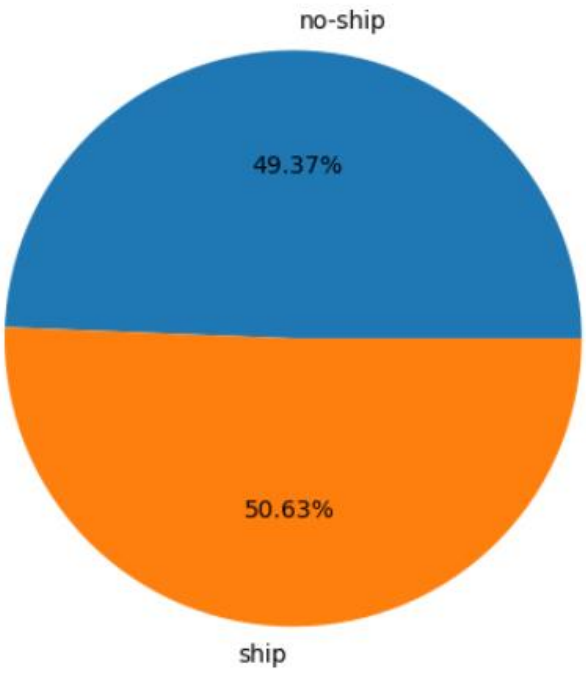


Рисунок 2 – Результат виведення аугментованих даних

Продемонстрований метод має проблеми під час навчання, адже як в тестову вибірку, так і в навчальну будуть потрапляти одні і ті ж зображення, що може призводити до завчання моделлю результату замість знаходження закономірностей, зменшити негативний вплив можна змінюючи дублі: додаванням шумів, обертанням чи віддзеркаленням, зміною кольорів.

Крім продемонстрованого вище метода досягти зменшення негативного впливу дисбалансу класів можна введенням вагових коефіцієнтів для кожного класу, таким чином в обраному датасеті зросте штраф за нерозпізнаний корабель, а нагорода за розпізнану відсутність корабля впаде, що нівелює вплив дисбалансу на результат навчання моделі, одначе потребує додаткових обчислень.

Отже, продемонстровані методи боротьби з дисбалансом класів дозволяють уникати негативного впливу на якість передбачення навченої моделі, що було продемонстровано на прикладі обраного датасету.

*Список використаних джерел*


1. *Survey on deep learning with class imbalance [Електронний ресурс]. Режим доступу до ресурсу: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0192-5*
2. *Japkowicz N. The class imbalance problem: Significance and strategies. In: In proceedings of the 2000 international conference on artificial intelligence (ICAI). 2000;111–7.*
3. *Ships in Satellite Imagery [Електронний ресурс] // Online Journal for Research and Education. – 2021. – Режим доступу до ресурсу: https://www.kaggle.com/datasets/rhammell/ships-in-satellite-imagery*


**УДК 004.01**

*Триконенко С.В., студент 2 курсу магістратури*
*спеціальності 122 «Комп'ютерні науки»*
*Бабаков Р.М., доцент,*
*доцент кафедри інформаційних технологій*

## РОЗРОБКА І ДОСЛІДЖЕННЯ МЕТОДІВ ПАРСИНГУ БІБЛІОГРАФІЧНИХ ДАНИХ

*Донецький національний університет імені Василя Стуса, м. Вінниця*

В Україні та в інших країнах світу, функціонують бібліотеки, які становлять основу культурної, наукової, інформаційної та освітньої інфраструктури. Серед них можна виділити публічні бібліотеки, які є центрами культурного життя суспільства. Головних завданням бібліотек є усунення