Issue 12, 2022

https://doi.org/10.23939/

УДК 004.8:612.7

АНАЛІЗ КАРКАСНИХ МЕРЕЖ ВИЯВЛЕННЯ ОЗНАК В МОДЕЛЯХ ГЛИБИННОГО НАВЧАННЯ

Павло Пукач

Національний університет "Львівська політехніка",
pavlopukach@gmail.com, ORCID 0000-0002-0488-6828

© Пукач П., 2023

У роботі проаналізовано та порівняно сучасні моделі глибинного навчання для задач класифікації зображень МРТ колінного суглоба. Проаналізовано сучасні глибинні архітектури комп'ютерного зору для виділення ознак із зображень МРТ. Такий аналіз використано для створення прикладних архітектур моделей машинного навчання. Вказані моделі орієнтовані на автоматизацію процесу діагностування травм коліна у медичних приладах та системах. Досліджено різні типи каркасних мереж виявлення ознак для архітектур машинного навчання, які здійснюють класифікацію зображень магнітно-резонансної томографії (МРТ) коліна. Результуючі моделі оцінено на наборі даних MRNet із обчисленням метрик F1 та К-Каппи Коена. Результати роботи показують, що метрика Каппа Коена важлива для оцінювання моделей на архітектурі MRNet, оскільки забезпечує глибше розуміння класифікаційних рішень кожної моделі.

Ключові слова: МРТ коліна; автоматизоване діагностування; MRNet; глибинне навчання; комп'ютерний зір.

Вступ

Магнітно-резонансна томографія (МРТ) є ефективним методом діагностування травм колінних суглобів. МРТ є найкращим способом візуалізації коліна для виявлення, аналізу можливої патології та скерування пацієнта на лікування [1]. МРТ дає змогу створювати послідовності зображень у трьох різних розрізах – аксіальному, корональному і сагітальному. МРТ є основним інструментом для лікарів-травматологів, рентгенологів опорно-рухового апарата. Останні статистичні прогнози свідчать про значний дефіцит експертів-рентгенологів, а також інших спеціалістів у галузі медицини [2]. У зв'язку з постійним підвищенням результативності МРТ інтерес до такого методу діагностування постійно зростає як серед науковців у галузі штучного інтелекту, які можуть досліджувати і застосовувати сучасні методи машинного навчання для таких завдань, так і серед лікарів, для яких основною цінністю таких технологій є спрощення завдань діагностування [3, 4].

Постановка проблеми

Під час побудови моделі бінарної класифікації використовують багато способів інтерпретації та оцінювання результатів на основі передбачень моделі. Добре відомі показники оцінювання, такі як площа під кривою робочих характеристик приймача (ROC-AUC), ґрунтуються на матриці помилок. Однак не всі показники добре працюють на незбалансованих наборах даних. Незбалансовані дані стосуються тих типів наборів даних, де у цільовому класі нерівномірний розподіл спостережень, тобто одна мітка класу має порівняно велику кількість спостережень, тоді як

інший клас незначну кількість спостережень порівняно із загальною кількістю. Набір даних MRNet, на якому ґрунтуються дослідження у цій статті, можна вважати незбалансованим набором даних. Із загалом 1370 досліджень МРТ колінного суглоба набір даних містить 1104, або 80,6 % аномальних досліджень. Кількість МРТ колінного суглоба, які вважаються нормальними, становить лише 19,4 %. Беручи це до уваги, результати оцінювання для моделей, описаних у статті, не можуть покладатися тільки на ROC-AUC або прості показники точності прогнозування, оскільки дисбаланс у формі відсутності нормального обстеження коліна є значним. Отже, з метою подальшого та точнішого оцінювання моделей машинного навчання ми також вводимо розрахунок оцінки Каппа Коена.

Аналіз останніх досліджень та публікацій

Серед усіх робіт, пов'язаних із застосуванням методів глибинного навчання для класу задач, описаних вище, було декілька спроб перетренувати оригінальну модель MRNet на сучасніших архітектурах комп'ютерного бачення [5, 6]. Також досі не зафіксовано жодної спроби задокументувати поступове вдосконалення точності передбачення MRNet із використанням новіших архітектур комп'ютерного бачення.

В оригінальній статті Дж. Коен [7] описує свою статистику Каппа як частку узгодженості після того, як випадкову узгодженість вилучено з розгляду. Коефіцієнт Каппа Коена (K) – це статистичний показник, який використовують для вимірювання надійності між оцінювачами для якісних елементів. Зазвичай вважають, що це надійніший показник, ніж простий розрахунок відсотка узгодженості, оскільки к враховує можливість випадкового збігу. Ця статистика часто використовується для перевірки надійності декількох оцінювачів. Важливість надійності оцінювача полягає у тому, що він відображає ступінь, до якого дані, зібрані в дослідженні, є правильним поданням вимірюваних змінних величин [8].

З погляду Коена очікується, що ми матимемо два незалежні оцінювачі або класифікатори, причому категорії під класифікацією будуть незалежними, взаємовиключними та вичерпними. Щоб досягти цих умов, ми інтерпретуємо перший оцінювач як нашу модель класифікації для діагнозу – розрив ПКС, розрив меніска або загальна аномалія. Другим оцінювачем є набір міток перевірки. Хоча, за термінологією Коена, не існує "правильних" чи "неправильних" суджень, у нашому випадку прогнози, які зробив другий оцінювач, вважаються "істинними".

Значення Каппа Коена (K) можна розрахувати так:

$$K = \frac{p_0 - p_e}{1 - p_e},$$

де p_0 – спостережувана точність або частка прогнозів, оцінки яких збігаються; p_1 – коефіцієнт прогнозів, для яких узгодження очікується випадково, або очікувана точність. Як і багато інших метрик оцінки для моделей класифікації, розрахунок Каппа Коена спирається на матрицю помилок. На відміну від обчислення загальної точності, Каппа Коена враховує дисбаланс у розподілі класів для набору даних перевірки.

Приймемо матрицю помилок для окремого класифікаційного завдання (ненормальний, розрив ACL, розрив меніска), як у табл. 1, де TP – True Positives (істинні позитивні результати); FP – False Positives (хибні позитивні результати); FN – False Negatives (хибні негативні результати); TN – True Negatives (істинні негативні результати). Позначивши через N загальну кількість спостережень, можемо розрахувати необхідні значення для статистики Каппа в нашому випадку за такими формулами:

$$\mathrm{p_0} = \frac{\mathrm{TP+TN}}{\mathrm{N}},$$

$$\mathrm{p_E} = \frac{\mathrm{R_{cond}+R_{norm}}}{\mathrm{N}},$$

$$\mathrm{R_{cond}} = \frac{\mathrm{(TP+FP)\cdot(TP+FN)}}{\mathrm{N}},$$

$$R_{norm} = \frac{(TN+FP)\cdot(TN+FN)}{N}.$$

Таблиця 1

Матриця помилок для моделі класифікації на основі MRNet

	Прогнозовані патологічні стани	Прогнозовані нормальні випадки
Істинні патологічні стани	TP	FP
Справжні нормальні випадки	FN	TN

Спостережувана точність p_0 – це кількість випадків, правильно класифікованих у матриці помилок. Це означає рівень узгодженості між правдивими та прогнозованими даними. $\mathrm{R_{cond}}$ і R_{norm} можна описати як коефіцієнти для заданої мітки порівняно із загальною кількістю спостережень. Це та частина, яка визначає дисбаланс набору даних перевірки.

Хоча інтерпретація показника Каппа може бути складнішою, ніж для традиційних оцінювальних метрик, оскільки вона розташована в інтервалі [−1, 1], вважаємо, що вона має потенціал для розкриття кращих моделей класифікації із вилученням фактора випадкового вгадування.

Формулювання цілей статті

Цілі роботи такі: 1) порівняльний аналіз, що ґрунтується на наборі даних MRNet, з обчисленням метрик (ROC-AUC) оцінки F1 та К-Каппи Коена у задачах виявлення розриву зв'язок (ACL), для виявлення аномалій коліна, зокрема розриву зв'язок та меніска; 2) визначення переваг застосування метрики Каппа Коена для оцінювання моделей на архітектурі MRNet.

Виклад основного матеріалу

Результуюча архітектура та процес навчання. У цій роботі створено та навчено кілька моделей класифікації на оригінальній архітектурі MRNet із використанням сучасніших магістральних мереж як екстракторів функцій для одного блока MRNet, ураховуючи VGG11, VGG16, Resnet та Efficientnet. Оригінальна версія AlexNet архітектури MRNet також навчена для порівняння результатів оцінювання новішої моделі. Стратегія оцінювання полягає у вимірюванні показників ефективності класифікації, таких як ROC-AUC, оцінка простої точності та оцінка *F1*. Крім того, оцінка Каппа Коена буде обчислена для кожної моделі та кожного типу діагнозу. Кожен блок навченої моделі відповідає вихідній архітектурі MRNet. Блок працює в одній площині МРТ і вчиться виділяти з цієї площини релевантні ознаки упродовж навчання. У всіх моделях, поданих у цій статті, єдиними різними шарами є рівень виділення ознак і різні вхідні розміри класифікатора, оскільки не всі магістралі видобувають вектор ознак однакових розмірів із послідовності вхідних зображень. Результуючу архітектуру блоків моделі, використану в цій роботі, наведено на рис. 1.

Зазначимо, що вхідні дані опрацьовуються різними архітектурами CV (магістралі), а адаптивне об'єднання середніх значень застосовується до вихідних даних магістралі для кожної частини МРТ. Згодом повністю підключений рівень класифікатора повертає ймовірність для конкретного діагнозу.

Кожна модель містить три навчальні блоки, які тренувалися виключно на аксіальному, корональному та сагітальному МРТ-зрізах відповідно. Підхід логістичної регресії (LR) використано для об'єднання результатів трьох окремих блоків MRNet для кожної площини. Це забезпечує останній рівень агрегованих прогнозів у результаті консенсусу між рішеннями, прийнятими кожним із блоків для кожної площини. Перетворення вхідного зображення було використано, щоб запобігти перенавчанню. Розпочиналося воно з навчання кожного з базових блоків MRNet на кожній площині МРТ. Навчальний процес реалізовано на платформі Google Cloud Platform, зокрема із використанням сервісу Vertex AI. Загалом було навчено 45 моделей у хмарі (три навчальні завдання для кожного

типу діагностики, помножені на три площини МРТ, помножені на кількість магістралей – п’ять). Кожне навчальне завдання прискорено за допомогою 2×NVIDIA Tesla P100 GPU.

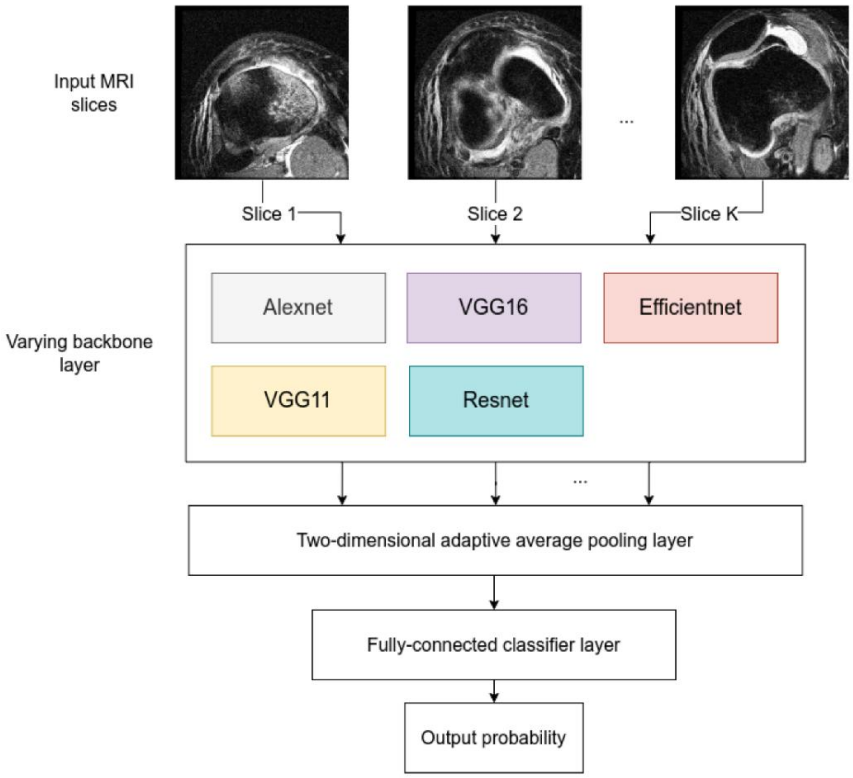


Рис. 1. Адаптована архітектура окремого блока MRNet

Кожен із блоків був навчений упродовж десяти епох, з оцінкою моделі на наборі даних перевірки після закінчення кожної епохи. Наприкінці кожної епохи робили знімок блока та завантажували його в Google Cloud Storage. Після тренувального процесу були відібрані лише блоки з найкращими оцінювальними балами. Потім отримані 45 блоків було зібрано з хмарного сховища та згруповано за типом магістралі. Після цього модель логістичної регресії було навчено на моделях для кожної площини, щоб зважити рішення щодо класифікації кожної з них відповідно до справжніх даних мітки перевірки. Навчання LR здійснювалося на локальній машині з використанням NVIDIA GeForce GTX 1080 Ti для пришвидшення процесу.

Після цього комбіновані прогнози моделі були зібрані в наборі даних перевірки. Відповідні вихідні дані моделі було збережено у файлах CSV для використання під час оцінювання моделі. Оцінювальні показники та графіки були побудовані на основі цих прогнозованих даних.

Загальна точність моделі. Першим очевидним і значущим показником оцінювання цього завдання класифікації є загальна точність моделі. У табл. 2 подано порівняння точності моделей для кожного діагнозу, а також середню точність для трьох діагнозів.

Таблиця 2

Порівняння точності моделей для різних діагнозів (поріг = 0,5)

Діагностика	Alexnet	VGG11	VGG16	Resnet	Efficientnet
	0,825	**0,858**	0,842	**0,858**	0,850
	0,683	0,792	**0,808**	0,583	0,550
	0,70	0,733	**0,750**	0,608	0,583
	0,738	0,752	**0,799**	0,683	0,661

Оскільки модель класифікації повертає ймовірності замість фактичних позитивних чи негативних міток, прогнози моделі були перетворені із порогом ймовірності 0,5. Усі ймовірності, що перевищують або дорівнюють 0,5, вважалися “позитивною” міткою. Максимальні значення виділено жирним шрифтом.

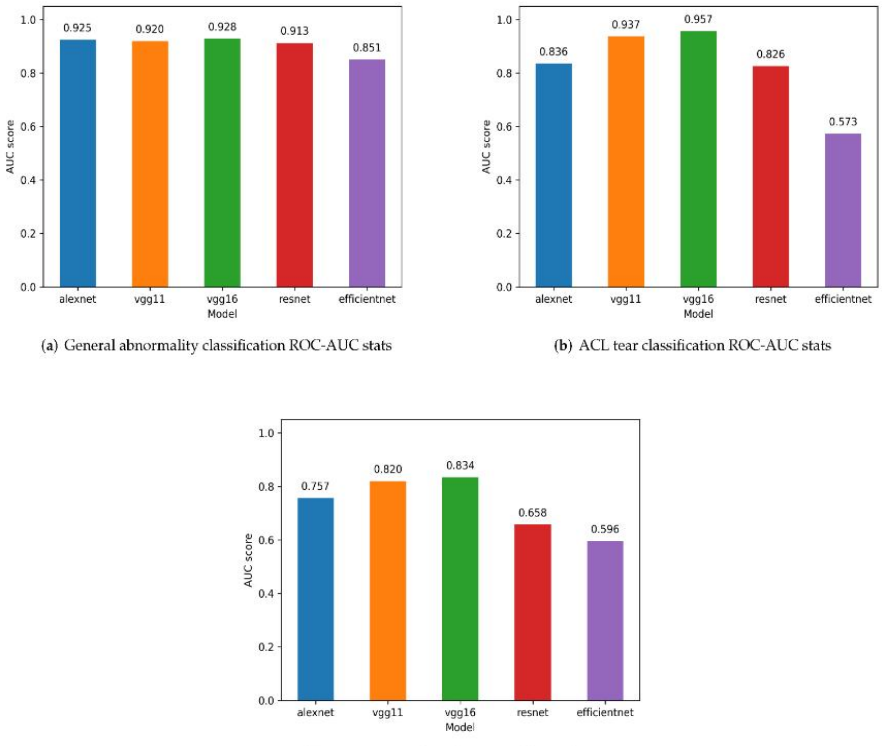
Оцінка F1. Іншим важливим показником оцінювання є оцінка *F1*, оскільки вона добре відповідає фактичному дисбалансу набору даних, збалансовує точність і запам’ятовування для позитивного класу, який домінує у цьому наборі даних. У табл. 3 наведено показник *F1*, розрахований для кожної навченої моделі та для кожного діагнозу. Вихідні прогнози перетворено так само, як і для загальної точності моделі, із використанням порогового значення 0,5.

Таблиця 3

***F1*-оцінка кожної моделі, для кожного діагнозу (порогове значення = 0,5)**

Діагностика	Alexnet	VGG11	VGG16	Resnet	Efficientnet
	0,900	**0,917**	0,909	**0,917**	0,912
	0,486	0,713	**0,753**	0,167	0,000
	0,673	0,628	**0,717**	0,299	0,324
	0,686	0,752	**0,793**	0,461	0,412

Результати ROC-AUC. Далі відображаються робочі характеристики приймача. На рис. 2 подано показники ROC-AUC для кожної моделі.



(a) General abnormality classification ROC-AUC stats

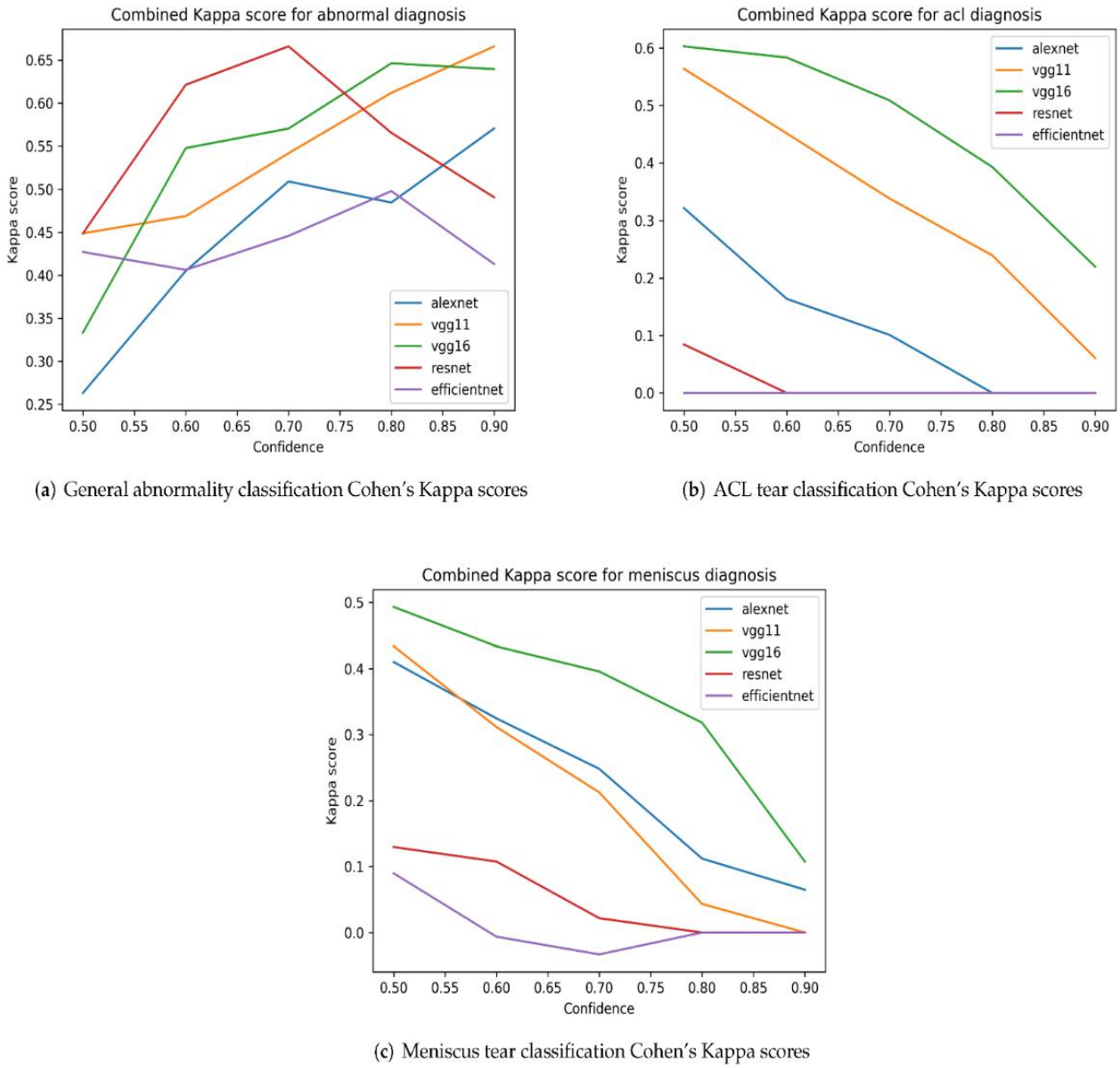
(b) ACL tear classification ROC-AUC stats

(c) Meniscus tear classification ROC-AUC stats

Рис. 2. Статистика ROC-AUC для кожної моделі та для кожного типу діагнозу:
a – класифікація загальної аномалії; b – класифікація розривів ACL; c – класифікація розривів меніска

Кожна панель відповідає певному типу діагнозу. Максимальні значення виділено жирним шрифтом.

Оцінки Каппа Коена. Вважаємо, що найзначущішим показником оцінки для цього дослідження є значення статистики Каппа Коена. Через незбалансованість набору даних MRNet цілком імовірно, що оцінка моделі в перспективі надійності між оцінювачами може надати більше корисної інформації, ніж традиційні оцінювальні бали. Знов-таки, для перетворення прогнозів моделі з ймовірностей на фактичні мітки були використані різні порогові значення в діапазоні від 0,5 до 0,9. Це дало змогу краще проілюструвати загальну поведінку певних моделей, не фіксуючись на одному значенні для порога. На рис. 3 наведено графік значень Каппа Коена для певних порогових значень, або значень довіри.



(a) General abnormality classification Cohen's Kappa scores

(b) ACL tear classification Cohen's Kappa scores

(c) Meniscus tear classification Cohen's Kappa scores

Рис. 3. Показники Каппа Коена для моделі та для кожного типу діагнозу: a – класифікація загальної аномалії; b – класифікація розривів ACL; c – класифікація розривів меніска

Продуктивність одного блока для кожного типу магістралі. Ще одним дуже важливим аспектом оцінювання цієї моделі є продуктивність кожної магістралі, перерахованої у цій статті, на площину. Ці результати відображено в табл. 4.

Таблиця 4

Оцінка ROC-AUC кожного хребта на площину МРТ

Хребет	Площина	Загальна аномалія	ACL	Розрив меніска
Alexnet	Аксіальна	0,845	0,738	**0,813**
	корональна	0,788	0,616	0,633
	сагітальна	**0,935**	0,834	0,707
VGG11	аксіальна	0,892	0,758	**0,811**
	корональна	0,732	0,926	0,762
	сагітальна	**0,917**	**0,903**	0,753
VGG16	аксіальна	**0,919**	**0,829**	0,759
	корональна	**0,840**	**0,950**	**0,808**
	сагітальна	0,909	0,883	**0,801**
Resnet	аксіальна	0,848	0,763	0,618
	корональна	0,477	0,462	0,659
	сагітальна	0,912	0,700	0,632

Найвищі бали оцінювання площини та для конкретного діагнозу виділені. Вони показують неопрацьований результат оцінювання класифікації ROC-AUC для кожного блока MRNet, навченого на певних магістралях. Рівень логістичної регресії не бере участі в цьому оцінюванні, тобто ця таблиця відображає єдину продуктивність цієї магістральної мережі на трьох зрізах МРТ для трьох завдань класифікації.

Висновки

Експериментально здійснено порівняння підвищення продуктивності встановлення діагнозу із використанням каркасної моделі Alexnet до магістралей VGG11 і VGG16 для класифікації оригінальних даних зображення MRNet [9]. Дослідження виявило, що продуктивність вищезгаданих моделей хребта різна в різних площинах МРТ. Щоб досягти найвищої точності класифікації, доцільно використовувати ансамбль логістичної регресії (LR) різних каркасних мереж виділення ознак: VGG16 – на корональному розрізі для усіх завдань класифікації; ACL – на аксіальному розрізі для виявлення аномальної форми коліна та розриву зв'язки; Alexnet – на сагітальному розрізі для виявлення аномальної форми коліна та на аксіальному розрізі для виявлення розриву меніска; VGG11 – на сагітальному розрізі для виявлення розриву зв'язки ACL. Підхід до оцінювання надійності та оцінювання продуктивності моделі у формі метрики Каппа Коена, що розрахована для кожного діагнозу та кожного типу каркасу виділення ознак, є значущою оцінкою, оскільки вона може надати глибше розуміння продуктивності моделі, ніж традиційні показники ROC-AUC. Результати роботи свідчать, що метрика Каппа Коена важлива для оцінювання моделей на архітектурі MRNet, оскільки забезпечує глибше розуміння класифікаційних рішень кожної моделі.

Список літератури

1. Nacey, N. C. (2017). Magnetic resonance imaging of the knee: An overview and update of conventional and state of the art imaging. *J. Magn. Reson. Imaging*, 45 (5), 1257–1275. https://doi.org/10.1002/jmri.25620.

2. IHS Markit Ltd (Prepared for the AAMC). The Complexities of Physician Supply and Demand: Projections from 2019 to 2034 AAMC (2021). https://www.aamc.org/media/54681/download.

3. Gore, J. C. (2020). Artificial intelligence in medical imaging. *J. Magn. Reson. Imaging*, 68, A1-A4. https://doi.org/10.1016/j.mri.2019.12.006.

4. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv, arXiv:1512.03385. https://doi.org/10.48550/arXiv.1512.03385.

5. Tsai, C., Kiryati, N., Konen, E., Eshed, I., & Mayer, A. Knee Injury Detection using MRI with Efficiently-Layered Network (ELNet). *In Proceedings of the Third Conference on Medical Imaging with Deep Learning*, Montreal,

QC, Canada, 6–8 July 2020, 121, 784–794.

6. Пукач П. (2022). Огляд та аналіз основних каркасних мереж виявлення ознак для класифікації зображень МРТ в моделях глибинного навчання. *Вісник Хмельницького національного університету. Технічні науки*, № 6 (315) [в друці].

7. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.*, 20, 37–46. https://doi.org/10.1177/001316446002000104.

8. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem. Med.*, 22, 276–282. https://doi.org/10.11613/BM.2012.031.

9. ImageNet Competition Leaderboard (2021). https://paperswithcode.com/sota/image-classification-on-imagenet.

References

1. Nacey, N. C. (2017). Magnetic resonance imaging of the knee: An overview and update of conventional and state of the art imaging. *J. Magn. Reson. Imaging*, 45 (5), 1257–1275. https://doi.org/10.1002/jmri.25620.

2. IHS Markit Ltd (Prepared for the AAMC). The Complexities of Physician Supply and Demand: Projections from 2019 to 2034 AAMC. (2021). https://www.aamc.org/media/54681/download.

3. Gore, J. C. (2020). Artificial intelligence in medical imaging. *J. Magn. Reson. Imaging*, 68, A1-A4. https://doi.org/10.1016/j.mri.2019.12.006.

4. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. arXiv, arXiv:1512.03385. https://doi.org/10.48550/arXiv.1512.03385.

5. Tsai, C., Kiryati, N., Konen, E., Eshed, I., & Mayer, A. Knee Injury Detection using MRI with Efficiently-Layered Network (ELNet). *In Proceedings of the Third Conference on Medical Imaging with Deep Learning*, Montreal, QC, Canada, 6–8 July 2020, 121, 784–794.

6. Pukach, P. (2022). Review and analysis of basic feature detection networks for classification of mri images in deep learning models. *Herald of Khmelnytskyi National University. Technical sciences*, No. 6 (315) [in press].

7. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.*, 20, 37–46. https://doi.org/10.1177/001316446002000104.

8. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem. Med.*, 22, 276–282. https://doi.org/10.11613/BM.2012.031.

9. ImageNet Competition Leaderboard (2021). https://paperswithcode.com/sota/image-classification-on-imagenet.

ANALYSIS OF FRAMEWORK NETWORKS FOR SIGN DETECTION IN DEEP LEARNING MODELS

Pavlo Pukach,
Lviv Polytechnic National University, pavlopukach@gmail.com,
ORCID 0000-0002-0488-6828

This paper analyzes and compares modern deep learning models for the classification of MRI images of the knee joint. An analysis of modern deep computer vision architectures for feature extraction from MRI images is presented. This analysis was used to create applied architectures of machine learning models. These models are aimed at automating the process of diagnosing knee injuries in medical devices and systems. This work is devoted to different types of feature detection framework networks for machine learning architectures that perform magnetic resonance imaging (MRI) image classification of the knee. The resulting models were evaluated on the MRNet validation dataset, calculating the metrics (ROC-AUC), prediction accuracy, F1 score, and Cohen's K-Kappa. The results of this work also show that Cohen's Kappa metric is important for evaluating models on the MRNet architecture because it provides a deeper understanding of the classification decisions of each model.

Key words: knee MRI; automated diagnosis; MRNet; deep learning; computer vision.