*Вестник Сыктывкарского университета.*
*Серия 1: Математика. Механика. Информатика.*
*Выпуск 1 (38). 2021*

**УДК 004.85: 551.734. DOI: 10.34130/1992-2752_2021_1_27**
**(470.4+574.1)**

# ПРОГРАММНО-ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ПАЛЕОПАЛИНОЛОГИЧЕСКОЙ ЗАДАЧИ

***В. В. Бабенко, Н. О. Котелина, О. П. Тельнова***

Геологическая информация имеет свою специфику. Несмотря на то что она считывается с природных объектов, результаты часто страдают субъективизмом и плохо воспроизводятся повторными опытами. Например, вторичное описание некоего геологического разреза обязательно будет отличаться. В силу этого высоко формализованные математические методы в геологии приживаются трудно. Наблюдавшийся в 70-е годы прошлого века всплеск интереса к математической геологии быстро затух. Напротив, значение информационных технологий практически для всех геологических наук стремительно возрастает. Геологический прогноз требует обобщения большого количества разнообразной фактографической информации и сведения ее в модели (в первую очередь, картографические), которые нужно визуально интерпретировать. В такой ситуации возможности современных баз данных и графический инструментарий компьютеров — настоящая находка.

В статье предлагаются перспективные варианты оптимизации исследований палеопалинологии (спор-пыльцевого анализа древних пород), являющейся одним из значимых разделов палеонтологии.

*Ключевые слова:* машинное обучение, deep learning, сверточная нейронная сеть, базы данных.

## 1. Введение

Палеопалинологическая задача — это определение относительного возраста горной породы по видовому составу фоссилизированных[1] микрообъектов: зерен спор и пыльцы древних растений. Аксиоматикой служат следующие эмпирически установленные тезисы:

- Массовость продуцирования спор и условия их переноса (воздушным и водным путем) и распространения позволяют говорить о статистических закономерностях распределения во вмещающих породах.
- Споры при фоссилизации сохраняют морфологические видовые признаки.
- Видовой состав спор, извлеченных (отмацерированных) из пород одного возраста сравнительно устойчив по площади (в пределах одного региона).
- Видовой состав спор значимо и закономерно меняется в различное геологическое время.
- Возраст образования вмещающих пород и синхронный возраст спор возрастает с увеличением глубины залегания пород[2].

Значимость такого определения возраста трудно переоценить — все геологические реконструкции и в конечном счете выводы, ориентированные на поиск полезных ископаемых, базируются на точной модели возрастных соотношений структурных элементов земной коры.

Формализация бизнес-процесса «Решение палеопалинологической задачи» показано на рис. 1. Основной результат (главный выход) процесса — это «O1 — Относительный возраст (стратиграфическое положение) вмещающей горной породы». Он должен явно соотноситься с миссией (стратегическими целями) организации. Миссией Института геологии КНЦ[3] является прогнозирование и поиск месторождений полезных ископаемых — точность определения геологического возраста является необходимой основой для моделирования геологического строения территории и последующего прогноза локализации месторождений.

[1] Окаменевших, замещенных вторичным минеральным веществом.

[2] Порядок чередования слоев пород может меняться под действием геологических процессов.

[3] Как и любой геологической организации.

Ключевым показателем эффективности процесса (KPI) является погрешность или вероятность ошибки определения возраста. Анализ метрик качества отдельных операций (табл. 1) позволяет оценить влияние последних на главный результат.

Детальный анализ вариативности управленческих факторов (см. рис. 1) и связанных с ними рисков выходит за рамки настоящей статьи. Отметим только, что ключевыми являются управления C1 — Личный опыт палинолога-эксперта и C5 — Диагностические атласы[4]. Первое полностью определяет качество основного результата и определяет его значительный субъективизм[5]. Проблема в том, что подготовка квалифицированного диагноста-палинолога — это длительная и дорогостоящая задача. Что касается атласов, то качество возможности поиска информации в них минимизированы и определяются, по сути, к той же персональной квалификации (предметной эрудиции) эксперта.

В статье описывается опыт оптимизации бизнес-процесса посредством экспериментального внедрения в практику информационных инструментов, улучшающих именно управления C1 и C5: личный опыт предложено улучшить с помощью использования интеллектуального алгоритма класса «deep learning», а информационную базу шаблонов (диагностические атласы) — путем формирования полноценной адаптивной базы данных.

## 2. База данных «Девонские споры»

Реляционная модель необходимой базы данных концептуально довольно проста (рис. 2). Она должна обеспечить накопление, хранение и возможность формирования многокритериальных выборок изображений девонских спор, точную привязку проб в трехмерном геологическом пространстве с сопутствующей информацией о вмещающих породах. А также позволять формировать и по необходимости менять важные концепты – палинокомплексы, являющиеся паттернами при принятии решения о возрасте вмещающих пород. Физическая реализация такой онтологии в среде MS SQL Server на рис. 3. Требуемый функционал программы, обеспечивающий эффективную работу эксперта-палинолога с такой базой данных, выявлен путем экспертного опроса и наблюдений

[4]Под этим термином здесь понимается весь доступный эксперту массив изображений классифицированных спор: публикации, результаты собственных исследований.

[5]Отметим, что субъективизм — это имманентная специфика геологического вывода.

Таблица 1

**Анализ пооперационных метрик качества бизнес-процесса «Решение палеопалинологической задачи»**

| Блок – Метрика | Комментарий | Единицы и диапазон измерений |
|---|---|---|
| A1 – Качество пробы | Проба породы отбирается с расчетом на содержание в ней значительного количества спор хорошей сохранности. Проверить этот критерий в поле невозможно. Применяются эмпирические правила: отбирать глину, аргиллиты, темно окрашенные, слабо метаморфизованные породы... | Эмпирическая вероятность (0..1) |
| A2 – Наличие спор в мацерате | На выходное количество спор влияет методика, которая является управлением C4 | Шкала: «пусто» - «единичные зерна» - «много» |
| A3 – Качество фотографий спор | Качество должно быть пригодным для диагностики. Определяется качеством аппаратуры (микроскоп и фотоаппарат) и опытом оператора | Шкала: «неприемлемо» - «хорошо» |
| A4 – Точность определения спор | Снижение уверенности в диагностических выводах возможно при наличии осложняющих факторов: плохой сохранности спор | Эмпирическая вероятность (0..1) |
| A5 – Репрезентативность палиноспектра | Наличие в палиноспектре пригодных к диагностике видов-индексов | Шкала: «неприемлемо» - «хорошо» |
| A6 – Вероятность ошибки определения возраста | Ошибка в определении возраста возможна всегда. Минимизировать ее позволяет личный опыт эксперта, который формализовать и алгоритмизировать сложно | Эмпирическая вероятность (0..1) |

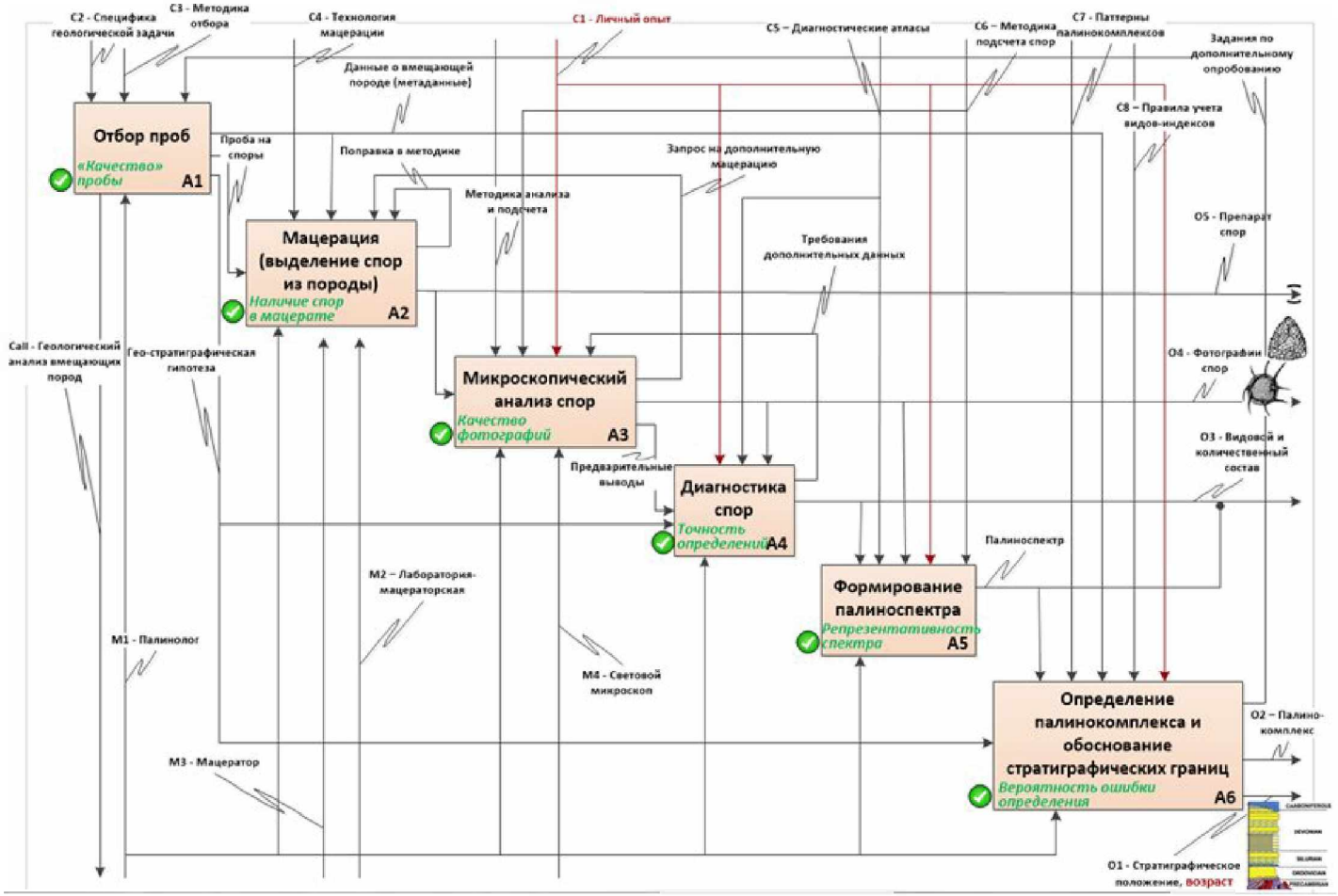


**Рис. 1.** SADT-модель [1] бизнес-процесса «Решение палеопалинологической задачи». Первая декомпозиция, «as is». Метрики качества отдельных операций помечены знаком $(\nu)$

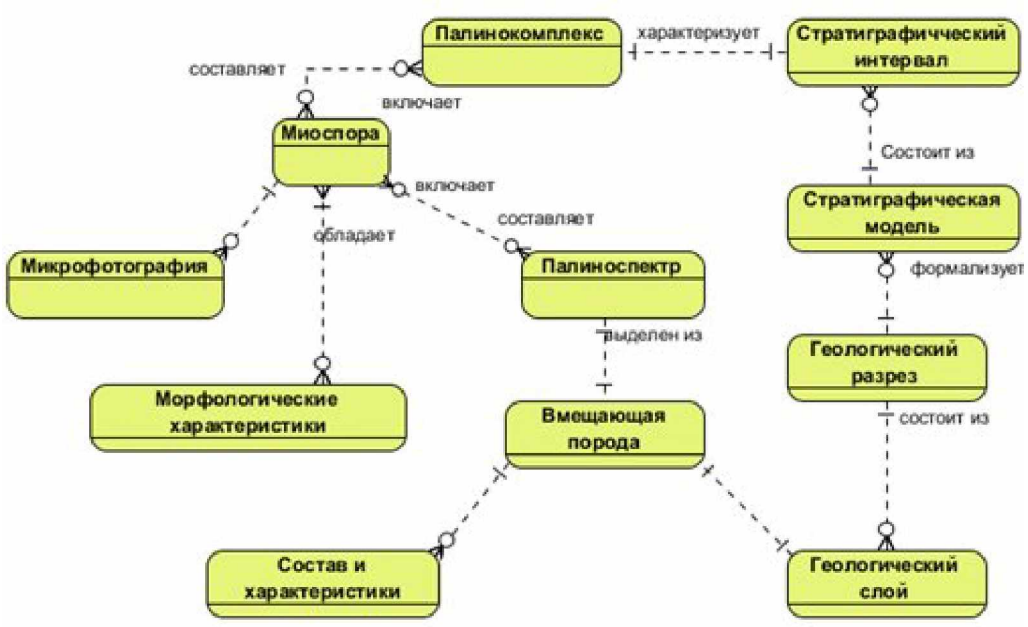


**Рис. 2.** Концептуальная ERD-модель базы данных «Девонские споры»

за реализацией конкретных реализаций процесса. В формате user stories требования выглядят так:

– Программа должна визуализировать точки отбора образцов (конкретные фотоизображения спор) на геологическом разрезе и выстраивать выборку изображений в линейной зависимости от глубины точки отбора.

– Должна обеспечиваться возможность выборки из базы данных изображений по запросу возраста, морфологических характеристик, принадлежности к палинокомплексу, конкретной родовой принадлежности и т.д.

– Все управление выборками должно подчиняться интуитивно понятным манипуляциям указателем мыши.

– Должна обеспечиваться возможность перемещения изображений по экранной форме с целью группирования их по различным критериям, а также масштабирование изображений. Весьма желательно использование алгоритма, облегчающего принятие решения о сходстве изображений исследуемых образцов с эталонами.

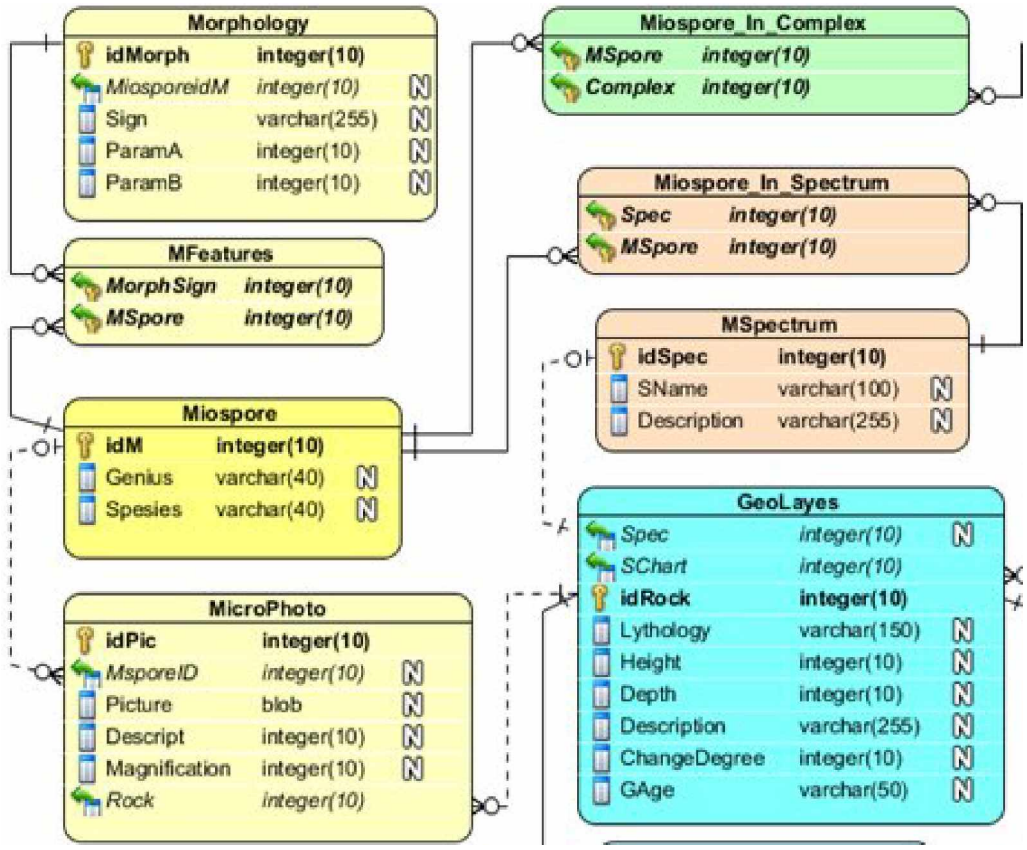


**Рис. 3.** Фрагмент физической ERD-модели базы данных «Девонские споры»

## 3. Использование алгоритма глубокого изучения для диагностики спор

Сложность диагностики видов индексов спор из древних пород — проблема исследовалась на объектах среднего и позднего девонского геологического возраста (интервал $\sim 350 - 400$ млн лет) — обусловлена плохой сохранностью спор, изменчивостью пространственной ориентации спор при захоронении, отсутствием строгой генетической классификации древних растений и четкого понимания закономерностей их морфологической эволюции [2]. Типичные микрофотографии сравнительно хорошего качества — именно с такими приходится иметь дело эксперту-палинологу — приведены на рис. 4. Как правило, диагностика проводится по морфологическим признакам (структура поверхности зерна, наличие внутренних структурных спецификаторов, соотношение внутренних оболочек и т. п.) [3]. Но алгоритмизировать выявление всех

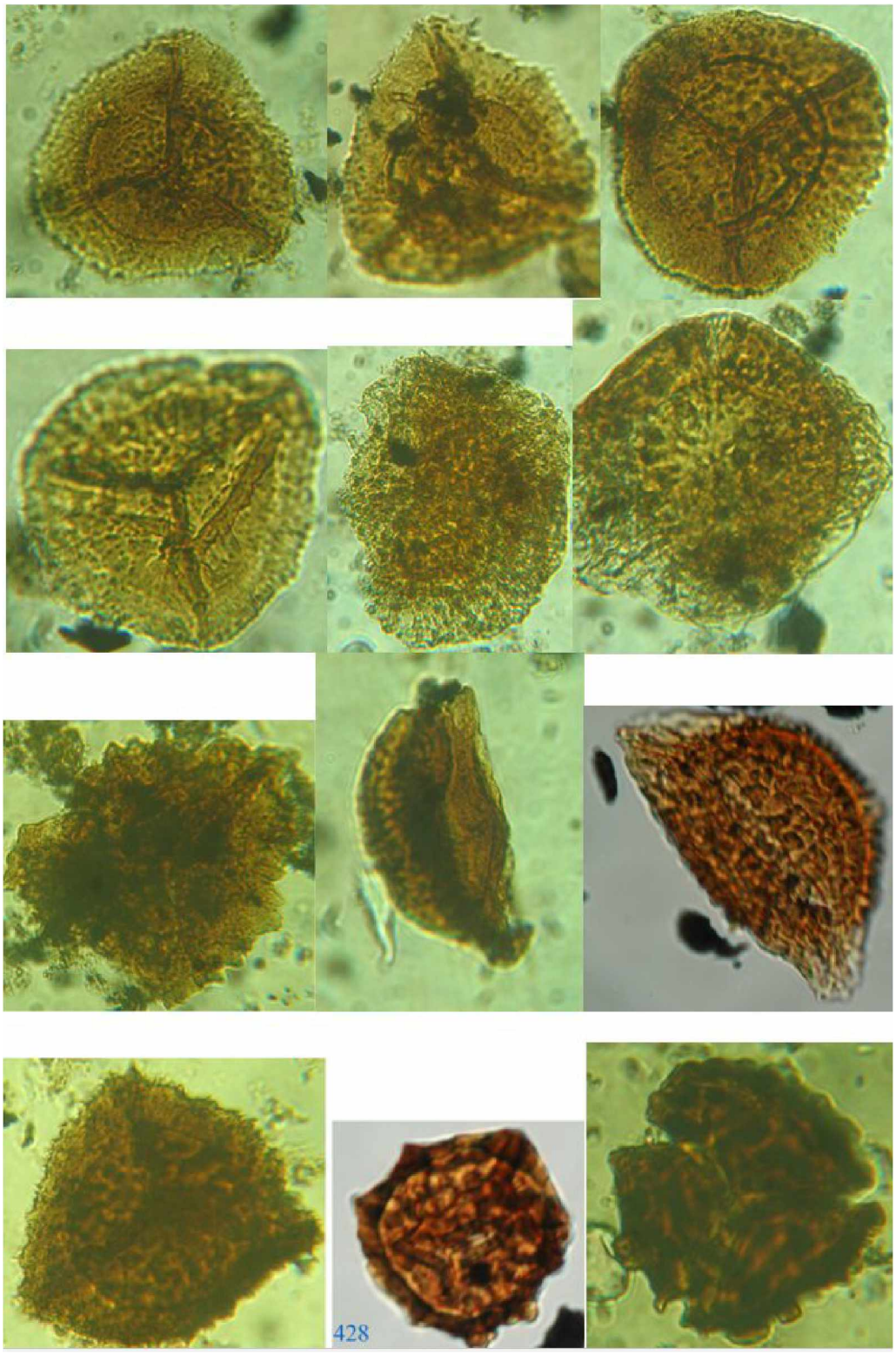


**Рис. 4.** Сводная фототаблица изображений девонских индексных спор из геологических разрезов Тимано-Печорской провинции [2]

этих признаков пока не удалось. Автоматическое распознавание морфологических признаков спор является отдельной задачей [4; 5], исследование которой запланировано нами на следующем этапе проекта. Была предпринята попытка эмулировать эмпирическое принятие классификационных решений путем использования алгоритмов глубокого изучения.

Идея во многом лежит на поверхности. Попытки глубокого изучения и автоматического распознавания спор предпринимались и ранее [4; 6; 7]. Правда работа выполнялась на современных или молодых плейстоценовых [7] спорах, что существенно упрощает решение – более качественные изображения, практически нет количественных ограничений при создании тренировочной выборки. При таких условиях удалось достичь качества в 97 % [6]. Все группы исследователей использовали метод сверточной нейронной сети (Convolutional Neural Network, CNN) [8]. Применялись техники трансферного обучения (transfer learning) и извлечения признаков (feature extraction) [6; 10].

Для оценки возможности автоматической классификации девонских спор нами также была использована последовательная модель Sequential() из Keras, высокоуровневого API для глубокого обучения, интегрированного в стандартный фреймворк Python [10]. Выбрана сверточная нейронная сеть (Convolutional Neural Network, CNN), известная своими преимуществами перед другими моделями, например перед многослойным перцептроном, при решении задач, связанных с распознаванием изображений [8; 11]. Архитектура построенной модели CNN является стандартной для распределения изображений по нескольким классам: это цепочка чередующихся слоев свертки Conv2D /подвыборки MaxPooling2D и несколько полносвязных слоев Dense с завершающим слоем с функцией активации softmax. После каждого сверточного слоя ко всем выходным значениям применяется функция активации ReLU.

Поскольку наша экспериментальная выборка достаточно ограничена[6], то нейронная сеть будет иметь склонность к переобучению. Для борьбы с переобучением нами используется техника dropout – выборочное отключение нейронов в процессе обучения [12]. Также проблему пе-

[6]Начальная экспериментальная выборка составила несколько десятков микрофотографий спор трех девонских видов индексов: Cristatisporites deliquescens, Membrabaculisporis radiatus и Archaeoperisaccus concinnus.

реобучения мы решаем при помощи генерации новых изображений, применяя ряд преобразований к исходному набору из обучающей выборки при помощи класса keras.preprocessing.image.ImageDataGenerator [13].

Для вероятностной классификации спор в качестве функции потерь использована перекрестная энтропия. Считая, что исходные данные распределены по классам равномерно, берем в качестве метрики оценки качества распознавания точность (accuracy).

## 4. Выводы

Опытное обучение проводилось на изображениях спор девонского возраста из разрезов Тимано-Печорской провинции, Республика Коми [2].

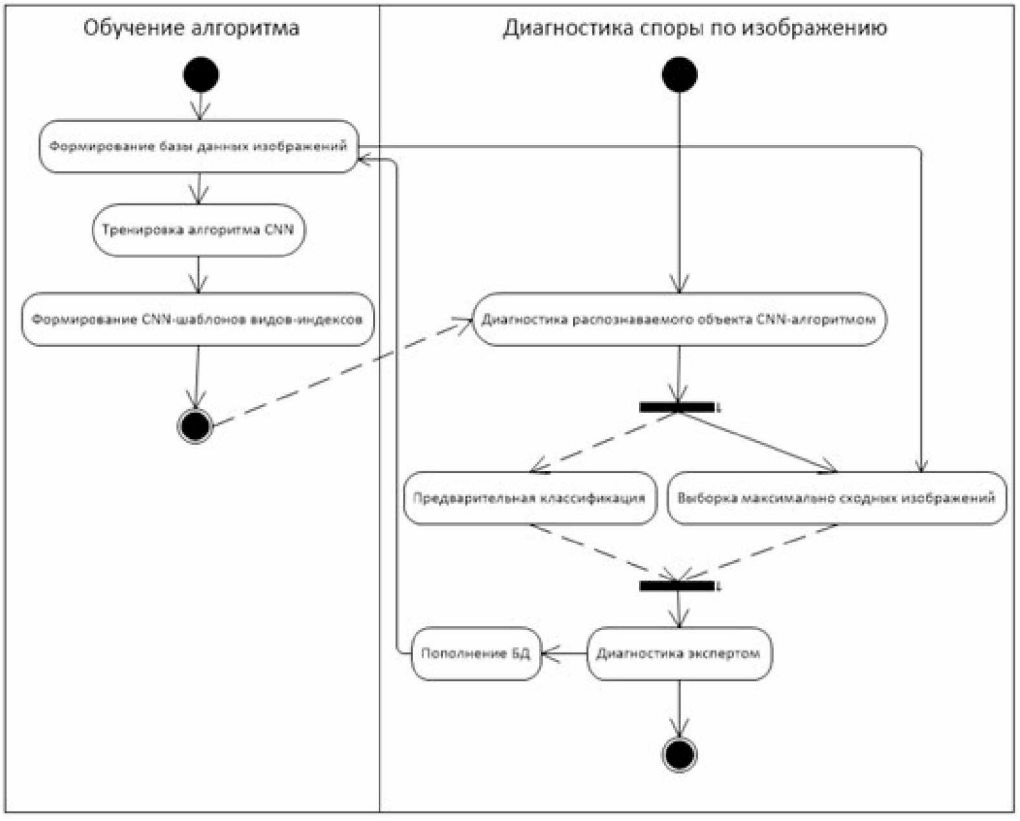


**Рис. 5.** Схема комбинированной диагностики видов-индексов «Коллаборация»

Полученное качество классификации мы оцениваем в 75 %. Это не позволяет пока говорить о полной автоматизации диагностики де-

вонских индексных спор, но показывает хорошие результаты в комбинированном алгоритме, который мы условно назвали «Коллаборация» (рис. 5) и в котором CNN-алгоритм выступает «предварительным экспертом» и фильтром при отборе из базы данных изображений, заслуживающих наибольшего внимания эксперта.

Экспериментальное использование алгоритма CNN и тестовая эксплуатация базы данных позволяет сделать следующие предварительные выводы:

Точность диагностики позволяет использовать технологию в качестве «интеллектуального помощника» эксперта.

Точность может быть повышена за счет увеличения тренировочной выборки.

Предполагается оценить возможность доработки алгоритма в части его адаптивности – при пополнении базы данных новыми образцами паттерны-эталоны должны пересчитываться.

Работы проводятся с частичной поддержкой гранта РФФИ № 20-05-00445.

## Список литературы

1. **Марка Д. А., МакГоуэн К.** Методология структурного анализа и проектирования SADT. М., 1993. 240 с.

2. **Tel'nova O. P., Shumilov I. Kh.** Middle–Upper Devonian Terrigenous Rocks of the Tsil'ma River Basin and Their Palynological Characteristics // *Stratigraphy and Geological Correlation, 2019. Vol. 27. No. 1. Pp. 27–50. DOI: https://doi.org/10.31857/S0869-592X27131-56.*

3. **Tel'nova O. P., Marshall J. E. A.** Devonian Spores of Kryshtofovichia africani Nikitin (Tracheophyta): Morphology and Ultrastructure // *Paleontological Journal, 2018. Vol. 52. No. 3. Pp. 342–349. © Pleiades Publishing, Ltd., 2018, published in Paleontologicheskii Zhurnal, 2018, No. 3, Pp. 119–124. ISSN 0031-0301. DOI: 10.1134/S0031030118030152.*

4. **Zhang Y., Fountain D. W., Hodgson R. M., Flenley J. R., Gunetileke S.** Towards automation of palynology: Pollen recognition using Gabor transforms and digital moments // *J. Quat. Sci. 19, 2004. Pp. 763–768.*

5. **Zhang Wang-Xiang, Zhao Ming-Ming, Fan Jun-Jun, Zhou Ting, Chen Yong-Xia,** & **Cao Fu-Liang.** Study on relationship between pollen exine ornamentation pattern and germplasm evolution in flowering crabapple. Sci., 2017. Rep. 7, 39759.

6. **Geus A. R. d., Barcelos C. A. Z., Batista M. A. and Silva S. F. d.** Large-scale Pollen Recognition with Deep Learning. 27-th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2019. Pp. 1–5. DOI: 10.23919/EUSIPCO.2019.8902735.

7. **C. Romero Ingrid, Kong Shu, C. Fowlkes Charless, Jaramillo Carlos, A. Urban Michael, Oboh-Ikuenobe Francisca, D'Apolito Carlos, W. Punyasena Surangi.** Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. Proceedings of the National Academy of Sciences Nov 2020, 117 (45), 28496-28505. DOI: 10.1073/pnas.2007324117.

8. **Huang G., Liu Z., Van Der Maaten L., Weinberger K. Q.** Densely connected convolutional networks. In: Conference on Computer Vision and Pattern Recognition, 2017. Vol. 1. Pp. 4700–4708.

9. **Do Chuong B., Ng Andrew Y.** Transfer learning for text classification. Neural Information Processing Systems Foundation, NIPS. 2005. URL: https://proceedings.neurips.cc/paper/2005/file/bf2fb7d1825a1df3ca308ad0bf48591e-Paper.pdf (дата обращения: 12.10.2020).

10. Keras: the Python deep learning API [Электронный ресурс] / официальный сайт Keras. URL: https:// keras.io (дата обращения: 12.10.2020).

11. Глубокое обучение: распознаем изображения с помощью сверточных сетей [Электронный ресурс] // Блог компании Wunder Fund, алгоритмы, машинное обучение. URL:

https://habr.com/ru/company/wunderfund/blog/314872/ (дата обращения: 12.10.2020).

12. **Gal Yarin, Ghahramani Zoubin.** Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning Proceedings of the 33-rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP. Vol. 48. URL: http://proceedings.mlr.press/v48/gal16.pdf (дата обращения: 12.10.2020).

13. The Keras Blog [Электронный ресурс]. URL: https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html (дата обращения: 12.10.2020).

14. **Sevillano V., Aznarte J. L.** Improving classification of pollen grain images of the POLEN23E dataset through three different applications of deep learning convolutional neural networks. PLoS ONE 13(9): e0201807. DOI: https://doi.org/10.1371/journal.pone.0201807. 2018.

**Summary**

**Babenko V. V.[1], Kotelina N. O.[1], Telnova O. P.[2]** Software and information support of the paleopalinological problem

Geological information has its own specifics. Despite the fact that it is read from natural objects, the results often suffer from subjectivity and are poorly reproduced by repeated experiments. For example, the secondary description of a geological section will necessarily differ. Because of this, highly formalized mathematical methods in geology are difficult to take root. The surge of interest in mathematical geology observed in the 70s of the last century quickly faded away. On the contrary, the importance of information technology for almost all geological sciences is growing rapidly. Geological forecast requires the generalization of a large amount of various factual information and its reduction in the model (first of all, cartographic), which must be visually interpreted. In such a situation, the capabilities of modern databases and graphical computer tools are a real boon.

The article proposes promising options for optimizing studies of paleopalinology (spore-pollen analysis of ancient rocks), which is one of the most important branches of paleontology.

*Keywords: machine learning, deep learning, convolutional neural network, databases.*

**References**

1. **Marca D. A., McGowan C. L.** *Metodologiya strukturnogo analiza i proyektirovaniya SADT* (Structured Analysis and Design Technique), Moscow, 1993, 240 p.

2. **Tel'nova O. P., Shumilov I. Kh.** Middle–Upper Devonian Terrigenous Rocks of the Tsil'ma River Basin and Their Palynological Characteristics, *Stratigraphy and Geological Correlation*, 2019, Vol. 27, No. 1, pp. 27–50. DOI: https://doi.org/10.31857/S0869-592X27131-56.

3. **Tel'nova O. P., Marshall J. E. A.** Devonian Spores of Kryshtofovichia africani Nikitin (Tracheophyta): Morphology and Ultrastructure, *Paleontological Journal*, 2018, Vol. 52, No. 3, pp. 342–349. © Pleiades Publishing, Ltd., 2018, published in Paleontologicheskii Zhurnal, 2018, No. 3, pp. 119–124. ISSN 0031-0301, DOI: 10.1134/S0031030118030152.

4. **Zhang Y., Fountain D. W., Hodgson R. M., Flenley J. R., Gunetileke S.** Towards automation of palynology: Pollen recognition using Gabor transforms and digital moments, *J. Quat.*, 2004, Sci. 19, pp. 763–768.

5. **Zhang Wang-Xiang, Zhao Ming-Ming, Fan Jun-Jun, Zhou Ting, Chen Yong-Xia,** & **Cao Fu-Liang.** Study on relationship between pollen exine ornamentation pattern and germplasm evolution in flowering crabapple., 2017, Sci. Rep. 7, 39759.

6. **Geus A. R. d., Barcelos C. A. Z., Batista M. A. and Silva S. F. d.** Large-scale Pollen Recognition with Deep Learning. 27-th European Signal Processing Conferencen (EUSIPCO), A Coruna, Spain, 2019, pp. 1–5. DOI: 10.23919/EUSIPCO.2019.8902735.

7. **C. Romero Ingrid, Kong Shu, C. Fowlkes Charless, Jaramillo Carlos, A. Urban Michael, Oboh-Ikuenobe Francisca , D'Apolito Carlos, W. Punyasena Surangi.** Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. Proceedings of the National Academy of Sciences Nov 2020, 117 (45), 28496-28505. DOI: 10.1073/pnas.2007324117.

8. **Huang G., Liu Z., Van Der Maaten L., Weinberger K. Q.** Densely connected convolutional networks. In: Conference on Computer Vision and Pattern Recognition. Vol. 1, pp. 4700–4708 (2017).

9. **Do Chuong B., Ng Andrew Y.** Transfer learning for text classification. Neural Information Processing Systems Foundation, NIPS. 2005. Available at: https://proceedings.neurips.cc/paper/2005/file/ bf2fb7d1825a1df3ca308ad0bf48591e-Paper.pdf (accessed: 12.10.2020).

10. Keras: the Python deep learning API [Electronic resource] / Keras official site. Available at: https:// keras.io (accessed: 12.10.2020).

11. Glubokoye obucheniye: raspoznavaniye izobrazheniy s pomoshch'yu svertochnykh setey (Deep learning: image recongition with convolutional neural networks) [Electronic resource] / Wunder Fund company blog, Algorithms, machine learning. Available at: https://habr.com/ru/company/wunderfund/blog/314872/ (accessed: 12.10.2020).

12. **Gal Yarin, Ghahramani Zoubin.** Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning Proceedings of the 33-rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP. Vol. 48. http://proceedings.mlr.press/v48/gal16.pdf.

13. The Keras Blog [Electronic resource] / Available at: https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html (accessed: 12.10.2020).

14. **Sevillano V., Aznarte J. L.** Improving classification of pollen grain images of the POLEN23E dataset through three different applications of deep learning convolutional neural networks. PLoS ONE 13(9): e0201807. DOI: https://doi.org/10.1371/journal.pone.0201807. 2018.

**Для цитирования:** Бабенко В. В., Котелина Н. О., Тельнова О. П. Программно-информационное обеспечение палеопалинологической задачи // *Вестник Сыктывкарского университета. Сер. 1: Математика. Механика. Информатика. 2021. Вып. 1 (38). С. 27–42. DOI: 10.34130/1992-2752_2021_1_27.*

**For citation:** Babenko V. V., Kotelina N. O., Telnova O. P. Software and information support of the paleopalinological problem, *Bulletin of Syktyvkar University. Series 1: Mathematics. Mechanics. Informatics*, 2021, 1 (38), pp. 27–42. DOI: 10.34130/1992-2752_2021_1_27.

*1 — СГУ им. Питирима Сорокина,*
*2 — Институт геологии Коми НЦ УрО РАН* *Поступила 01.03.2021*