

**Разработка метода автоматической идентификации границ и заголовков таблиц в Excel**

**Басилаев Д.И. (ИТМО), Першинов А.В. (ИТМО), Сапельникова К. С. (ИТМО),**

**Научный руководитель – кандидат технических наук, доцент Насонов Д. А.**

**(ИТМО), инженер, аспирант Ковальчук М. А.**

**(ИТМО)**

**Введение.** В современном мире, где объемы данных постоянно растут, важность эффективной и точной обработки таблиц в Excel становится все более очевидной. Сложность и разнообразие структур данных в таблицах требуют разработки продвинутых методов для их автоматического распознавания и обработки. Традиционные подходы к идентификации границ и заголовков таблиц часто сталкиваются с ограничениями, связанными с необходимостью ручной настройки и недостаточной универсальностью для обработки сложных и неоднородных данных. В этом контексте появляется потребность в разработке более мощных и адаптивных алгоритмов, способных автоматически определять структуру таблиц, распознавать заголовки и адаптироваться к различным форматам данных.

**Основная часть.** Исследование фокусируется на разработке метода, который интегрирует последние достижения в области машинного обучения и компьютерного зрения для решения задачи идентификации границ и заголовков таблиц в Excel. Основой метода является использование сверточных нейронных сетей (CNN), которые обучаются на большом объеме размеченных данных для выявления шаблонов и закономерностей, характерных для структур таблиц [1]. Этот подход позволяет не только точно определять границы таблиц, но и распознавать заголовки столбцов [2], что является ключевым для дальнейшей обработки и анализа данных. Кроме того, в методе используются эвристические подходы, которые, работая в комбинации с алгоритмами машинного обучения, обеспечивают более гибкое и многоаспектное решение проблемы, позволяя системе адаптироваться к разнообразным стилям и структурам данных, встречающимся в реальных документах Excel. Эта гибридная методология значительно повышает эффективность идентификации, обеспечивая высокую точность и надежность при обработке данных. Разработка такого метода требует тщательного анализа существующих подходов, выявления их недостатков и преимуществ, а также экспериментальной проверки предложенного решения на реальных данных.

Этот подход позволит не только существенно упростить и автоматизировать работу с таблицами в Excel, но и повысить точность и надежность обработки данных, что имеет важное значение для аналитической работы, исследований и принятия обоснованных решений на основе данных.

**Выводы.** Разработанный метод позволяет значительно повысить точность и скорость обработки данных в Excel, открывая новые возможности для аналитической работы и исследований в различных областях.

**Список использованных источников:**

1. Dong, H., Liu, S., Han, S., Fu, Z., and Zhang, D. Tablesense: Spreadsheet table detection with convolutional neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence (2019), vol. 33, pp. 69—76
2. Wang, Z., Dong, H., Jia, R., Li, J., Fu, Z., Han, S., & Zhang, D. (2021). Tuta: Tree-based transformers for generally structured table pre-training. In Proceedings of 27th ACM SIGKDD conference on knowledge discovery & data mining (pp. 1780–1790).