**УДК 004.89**

# РЕАЛИЗАЦИЯ МЕХАНИЗМОВ ЗАЩИТЫ ДАННЫХ В ДОВЕРЕННЫХ ML-БИБЛИОТЕКАХ

Киселёва С.А., студент гр. 23.М02-мкн, I курс
Санкт-Петербургский государственный университет, г. Екатеринбург

Современное развитие технологий машинного обучения сопровождается ростом объемов обрабатываемых данных, многие из которых содержат чувствительную или конфиденциальную информацию. В условиях усиления нормативных требований к защите персональных данных и угроз со стороны злоумышленников обеспечение безопасности и конфиденциальности информации становится ключевой задачей при разработке доверенных библиотек машинного обучения. Реализация эффективных механизмов защиты данных в таких библиотеках позволяет не только снизить риски утечек и несанкционированного доступа, но и повысить уровень доверия пользователей и регуляторов к ИИ-системам. В настоящей статье рассматриваются основные подходы и технологии, применяемые для защиты данных в рамках доверенных ML-библиотек [1].

Одним из фундаментальных механизмов является внедрение строгих политик контроля доступа, основанных на ролевой модели или принципах минимальных прав. Это позволяет ограничить доступ к данным и функциональности библиотеки только авторизованным субъектам, а также регламентировать действия внутри системы на основе контекстных условий. Важным аспектом является реализация механизмов аутентификации и авторизации, интегрированных с корпоративными системами управления идентификацией, что обеспечивает централизованный и безопасный контроль доступа.

Другой ключевой компонент – применение криптографических методов защиты данных. В доверенных ML-библиотеках всё чаще используются технологии дифференциальной приватности, позволяющие добавлять контролируемый шум к обучающим данным или градиентам, тем самым защищая информацию о конкретных записях в выборке без существенной потери качества модели. Кроме того, используются методы гомоморфного шифрования, позволяющие выполнять вычисления над зашифрованными данными, что существенно снижает риски раскрытия конфиденциальной информации во время обучения или инференса [2].

Важную роль играет реализация средств безопасного распределённого обучения, включая федеративные протоколы, при которых данные не покидают локальных устройств или серверов, а обмен информацией происходит исключительно в виде агрегированных обновлений модели. В таких сценариях библиотеки должны обеспечивать гарантии целостности и подлинности передаваемых данных, а также защищать коммуникационные каналы с помощью современных протоколов шифрования и аутентификации.

Для повышения надёжности защиты используются методы мониторинга и аудита действий с данными. Встроенные журналы фиксируют все операции с конфиденциальной информацией, что позволяет в режиме реального времени обнаруживать подозрительную активность, проводить расследования инцидентов и формировать отчёты для соответствия нормативным требованиям. Такие механизмы интегрируются с системами управления информационной безопасностью и помогают обеспечивать комплексный подход к защите данных [3].

Особое внимание уделяется защите данных на уровне памяти и вычислительных процессов. В доверенных библиотеках реализуются механизмы изоляции вычислений, предотвращающие несанкционированный доступ к данным во время выполнения, включая использование доверенных вычислительных сред (Trusted Execution Environments). Это позволяет минимизировать риски атак, направленных на извлечение информации через уязвимости операционной системы или аппаратного обеспечения [4].

Кроме технических решений, важным аспектом является соблюдение нормативных и этических стандартов в области защиты данных. Доверенные ML-библиотеки разрабатываются с учётом требований таких регламентов, как GDPR, HIPAA и других региональных законов, что требует реализации механизмов управления согласием пользователей, права на удаление данных и прозрачности процессов обработки информации.

Таким образом, реализация механизмов защиты данных в доверенных библиотеках машинного обучения представляет собой комплексную задачу, включающую в себя технические, организационные и нормативные компоненты. Эффективное сочетание криптографических методов, контроля доступа, безопасных протоколов обучения и мониторинга позволяет создавать надёжные и прозрачные инструменты для разработки и эксплуатации ИИ-систем. В перспективе дальнейшее развитие этих механизмов станет одним из ключевых факторов успешного и этически ответственного внедрения машинного обучения в критически важные сферы деятельности [5].

**Список литературы:**

1. Shokri, R. Membership Inference Attacks against Machine Learning Models / R. Shokri, M. Stronati, C. Song, V. Shmatikov // IEEE Symposium on Security and Privacy. – 2017. – P. 3–18. – DOI 10.1109/SP.2017.41.

2. Fredrikson, M. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures / M. Fredrikson, S. Jha, T. Ristenpart // Proceedings of the ACM CCS. – 2015. – P. 1322–1333. – DOI 10.1145/2810103.2813677.

3. Abadi, M. Deep Learning with Differential Privacy / M. Abadi, A. Chu, I. Goodfellow [et al.] // Proceedings of the ACM CCS. – 2016. – P. 308–318. – DOI 10.1145/2976749.2978318.

4. Bonawitz, K. Practical Secure Aggregation for Privacy-Preserving Machine Learning / K. Bonawitz, V. Ivanov, B. Kreuter [et al.] // Proceedings of the ACM CCS. – 2017. – P. 1175–1191. – DOI 10.1145/3133956.3133982.

5. Gentry, C. A Fully Homomorphic Encryption Scheme : dissertation / C. Gentry. – Stanford : Stanford University, 2009. – 209 p.