УДК 004.056, 004.8

РАЗРАБОТКА АНАЛИТИЧЕСКОГО ПРОГРАММНОГО КОМПЛЕКСА ПО ДЕТЕКТИРОВАНИЮ СТЕГАНОГРАФИЧЕСКИХ СООБЩЕНИЙ В ГРАФИЧЕСКИХ ИЗОБРАЖЕНИЯХ С ПРИМЕНЕНИЕМ ОБЪЯСНИТЕЛЬНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Скалдин Даниил Дмитриевич[1], Шевченко Алексей Валерьевич[2]

[1]Старший специалист отдела информационной безопасности;
ФКУ "Налог-Сервис" ФНС России по ЦОД в г. Дубна;
141981, Московская обл., г. Дубна, ул. Технологическая, 2;
e-mail: danya.skaldin@mail.ru.

[2]Аспирант;
Государственный университет «Дубна»;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: leviathan0909@gmail.com.

Работа посвящена разработке программного аналитического комплекса для обнаружения стеганографических сообщений в графических изображениях с использованием методов объяснительного искусственного интеллекта. Основная цель исследования — повышение эффективности стегоанализа за счёт применения технологий объяснимого ИИ, которые позволяют не только выявлять скрытую информацию, но и анализировать причины и обоснования полученных результатов. В рамках работы рассматриваются методы сокрытия данных, современные подходы к их обнаружению, а также реализуется система классификации изображений с возможным скрытым встраиванием, основанная на технологиях искусственного интеллекта, модулях объяснимого ИИ и собственном наборе данных. Разработанный комплекс может быть полезен специалистам в области информационной безопасности, цифровой экспертизы и другим заинтересованным пользователям.

Ключевые слова: обнаружение стеганографии, объяснительный искусственный интеллект (*XAI*), информационная безопасность, машинное обучение, собственный набор данных.

Для цитирования:

Скалдин Д. Д., Шевченко А. В. Разработка аналитического программного комплекса по детектированию стеганографических сообщений в графических изображениях с применением объяснительного искусственного интеллекта // Системный анализ в науке и образовании: сетевое научное издание. 2025. № 4. С. 9-23. EDN: KRFRGW. URL: https://sanse.ru/index.php/sanse/article/view/674.

DEVELOPMENT OF AN ANALYTICAL SOFTWARE PACKAGE FOR DETECTING STEGANOGRAPHIC MESSAGES IN GRAPHIC IMAGES USING EXPLANATORY ARTIFICIAL INTELLIGENCE

Skaldin Daniil D.[1], Shevchenko Alexey V.[2]

[1]Senior Specialist of the Information Security Department;
FPT "Nalog-Service" of FTS for DPC in Dubna;
2 Technologicheskaya Universitetskaya Str., Dubna, Moscow region, 141981, Russia;
e-mail: danya,skaldin@mail.ru.

[2]Graduate student;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;

e-mail: leviathan0909@gmail.com.

The work is devoted to the development of an analytical software package for detecting steganographic messages in graphic images using explanatory artificial intelligence methods. The main purpose of the study is to increase the effectiveness of stepanalysis through the use of explicable AI technologies that allow not only to identify hidden information, but also to analyze the reasons and justifications for the results obtained. The paper examines methods of data concealment, modern approaches to their detection, and implements an image classification system with possible hidden embedding based on artificial intelligence technologies, explicable AI modules, and its own dataset. The developed complex can be useful to specialists in the field of information security, digital expertise and other interested users.

Keywords: steganography detection, Explanatory artificial intelligence (XAI), information security, machine learning, proprietary dataset.

For citation:
Skaldin D. D., Shevchenko A. V. Development of an analytical software package for detecting steganographic messages in graphic images using explanatory artificial intelligence. *System analysis in science and education*, 2025;(4):9-23(in Russ). EDN: KRFRGW. Available from: https://sanse.ru/index.php/sanse/article/view/674.

Введение

В современном информационном обществе проблема защиты данных приобретает все большую актуальность. Один из аспектов информационной безопасности связан с обнаружением скрытых каналов передачи информации, в частности, стеганографических сообщений в цифровых изображениях.

Данная работа посвящена разработке аналитического программного комплекса, способного эффективно обнаруживать стеганографические сообщения в изображениях формата *JPG* и интерпретировать для пользователя результаты анализа. Комплекс объединяет в себе передовые методы цифровой обработки изображений, машинного обучения и технологии объяснительного искусственного интеллекта, что позволяет достичь высокой точности детектирования при сохранении прозрачности процесса анализа и тем самым решить проблему существующих систем, работающих по принципу «чёрного ящика».

Актуальность темы исследования обусловлена растущей потребностью в эффективных инструментах противодействия скрытым каналам передачи информации, которые могут использоваться для незаконной деятельности, включая промышленный шпионаж, распространение запрещенного контента и координацию действий злоумышленников. Разработка программного комплекса, сочетающего высокую точность детектирования с возможностью объяснения полученных результатов, представляет собой важный шаг в развитии средств обеспечения информационной безопасности.

1. Цифровая стеганография

1.1. Определение и сущность цифровой стеганографии

Сохранение конфиденциальности данных представляет собой одно из наиболее значимых вызовов современного информационного общества. Однако потребность в скрытой передаче информации возникла задолго до появления цифровых технологий и уходит корнями в глубокую древность. Уже в 5 веке д.н.э. в трудах Геродота упоминается метод передачи тайных сообщений: греческий тиран Гистий выбривал голову раба, писал на коже послание, а после отрастания волос отправлял его к адресату. Стороннему наблюдателю такой посланник казался обычным человеком, не вызывающим подозрений. Спустя много веков появились невидимые чернила, микротекст и другие способы скрытой передачи сведений, основная цель которых заключалась в том, чтобы сохранить тайну самого факта коммуникации [1].

С развитием цифровых технологий практика сокрытия информации трансформировалась в отдельное направление — цифровую стеганографию. Несмотря на активное распространение

соответствующих методов, единое и строго установленное определение данного термина в научной литературе отсутствует. На практике цифровую стеганографию трактуют как совокупность технических и алгоритмических средств, позволяющих встраивать скрытые сообщения в цифровые объекты (изображения, аудио- и видеофайлы, текстовые документы, сетевой трафик и др.) без заметных изменений их структуры и визуального или статистического восприятия. Главной задачей таких методов является сокрытие самого факта передачи информации. Однако такая незаметность делает цифровую стеганографию привлекательной не только для легитимных целей — например, защиты авторских прав или обеспечения конфиденциальности, — но и для злоумышленников, стремящихся обходить системы мониторинга и передавать данные скрытным образом.

1.2. Классификация цифровой стеганографии

Цифровая стеганография — это сложная, многогранная дисциплина, которую можно классифицировать по следующим признакам (см. рис. 1):

Рис. 1. Классификация цифровой стеганографии

1. *Классификация по типу контейнеров (цифровых объектов (файл или поток), в которые внедряется скрытая информация)* [2-3]*:*

   - Потоковые контейнеры — это данные, передающиеся непрерывным потоком (например сетевой трафик или аудио- и видеопотоки). В рамках потоковой стеганографии применяются скрытая передача данных (незаметное внедрение информации в параметры передачи) и цифровые водяные знаки (устойчивые, незаметные метки, встроенные в поток или файл, используемые для доказательства авторства или защиты прав).

   - Фиксированные контейнеры — это статические файлы, в которых структура остаётся неизменной после внедрения. В роли таких контейнеров могут быть изображения (информация внедряется, например, в биты, отвечающие за цвет пикселей), аудио и видео (скрытие информации в звуковых или видеокадрах), заголовки и метаданные (специальные служебные поля файлов, используемые для внедрения вспомогательной информации).

2. *Классификация по стойкости (описывает, насколько внедрённая информация сохраняется при различных воздействиях по типу сжатия или изменения размера на контейнер):*

   - Робастные методы (*robust*) — устойчивы к изменению формата, сжатию, повороту, фильтрации и другим обработкам. Часто применяются в цифровых водяных знаках, при необходимости сохранения встроенной информации при любых манипуляциях с файлом.

   - Хрупкие методы (*fragile*) — теряют встроенную информацию при малейшем изменении контейнера. Применяются для контроля целостности.

   - Полухрупкие методы (*semi-fragile*) — устойчивы к незначительным случайным изменениям, но уязвимы к злонамеренному вмешательству. Используются, например, в системах цифровой сертификации или контроля доступа.

3. *Классификация по степени скрытности (отражает, насколько метод скрытия зависит от секретности алгоритма):*

   - Закрытые методы — алгоритмы сокрытия не публикуются. Только создатели знают, как работает система. Если метод становится известен, вся система теряет безопасность.

   - Полузакрытые методы — часть алгоритма открыта, часть скрыта. Это позволяет комбинировать безопасность с практичностью. Широко применяются в коммерческих решениях.

   - Открытые методы — алгоритмы полностью документированы. Безопасность обеспечивается исключительно секретностью ключа.

4. *Классификация по типы детекторов (инструментов или алгоритмов, применяемых для выявления или извлечения скрытых данных):*

   - Декодер цифровых водяных знаков — извлекает встроенные метки (водяные знаки) при наличии ключа и информации о методе внедрения. Часто используется для подтверждения авторства.

   - Вероятностный детектор — использует статистические модели и методы машинного обучения. Он не извлекает сообщение напрямую, но определяет с высокой вероятностью, есть ли в контейнере скрытая информация.

   - Жёсткий детектор — выдаёт бинарный ответ: есть ли скрытая информация или нет. Необходим в условиях высокой критичности, например, при защите государственных информационных систем.

5. *Классификация по пространству встраивания данных:*

   - Вложение в исходный сигнал — данные внедряются непосредственно в содержимое файла (пиксели изображения (наименьшие логические элементы двумерного цифрового изображения в растровой графике), отсчёты звукового сигнала, символы текста). Пример — метод *LSB* (*Least Significant Bit*), при котором изменяются наименее значимые биты цвета изображения. Такой подход прост, но уязвим к любым преобразованиям файла.

   - Вложение в преобразованную область — информация внедряется после применения математического преобразования к исходному сигналу. Наиболее популярны:

   - *DCT* (дискретное косинусное преобразование) — применяется в *JPEG*-сжатии;

   - *DWT* (вейвлет-преобразование);

   - *DFT* (дискретное преобразование Фурье) [4].

Внедрение происходит в коэффициенты преобразования. Эти методы значительно устойчивее к сжатию и фильтрации. Известным примером является метод Коха–Жао, при котором данные внедряются в среднечастотные DCT-коэффициенты блоков JPEG-изображений. Такой подход обеспечивает хорошую маскировку и стойкость.

1.3. Области применения цифровой стеганографии

Цифровая стеганография применяется в разных сферах, где важна защита, скрытность или аутентичность данных:

- Защита авторских прав — цифровые водяные знаки позволяют доказать авторство и предотвратить незаконное распространение цифрового контента. Такие метки могут использоваться в судебных разбирательствах.
- Обеспечение конфиденциальности — сокрытие данных в изображениях или аудио позволяет передавать информацию без использования зашифрованных файлов, которые легко распознать.
- Контроль целостности — хрупкие методы позволяют определить, подвергался ли файл изменениям. Если встроенное сообщение повреждено, значит файл был изменён.

1.4. Методы стеганографического скрытия информации в графических изображениях

Графические изображения являются одними из наиболее распространённых контейнеров для цифровой стеганографии. Это связано с их визуальной избыточностью, популярностью в коммуникации и поддержкой различных форматов, позволяющих внедрять информацию как в пиксельное представление, так и в трансформированные частотные компоненты.

Методы сокрытия информации в изображениях можно классифицировать по области внедрения данных, принципам воздействия на содержимое и способу кодирования сообщений. Выделяют два основных класса: **пространственные** методы и **частотные.**

Пространственные методы основываются на непосредственном изменении значений пикселей изображения в его оригинальном (пространственном) представлении. Это означает, что скрытая информация внедряется напрямую в цветовые компоненты, не подвергая изображение каким-либо математическим преобразованиям.

Ключевая особенность пространственных методов заключается в высокой простоте реализации, значительной вместимости и высокой чувствительности к преобразованиям изображения.

Примеры подходов:

- Метод наименее значащего бита (*LSB*) — внедрение информации в младшие биты каналов *RGB* (цветовая модель, которая формирует цвета с помощью трёх каналов: красного, зелёного и синего), практически без влияния на визуальное восприятие;
- *Edge-based embedding* — внедрение данных в области изображения с резкими переходами (границами), чтобы минимизировать визуальные искажения;
- *PVD* (*Pixel Value Differencing*) — метод, использующий разности между соседними пикселями для кодирования информации.

Данные подходы имеют плюсы за счёт своей простоты реализации, высокой скорости и большой вместимости данных, а минусы заключаются в уязвимости к сжатию, низкой устойчивости к фильтрации и изменению размера, а также возможности детектирования с помощью гистограммного или визуального анализа.

Пространственные методы часто используются для несжатых форматов изображений (*BMP, PNG*), где сохраняется точная структура пикселей. Однако в *JPEG*-изображениях, где применяется сжатие с потерями, такие методы непрактичны — они неустойчивы и могут быть полностью разрушены при сохранении.

Частотные методы в отличие от пространственных используют предварительное математическое преобразование изображения перед внедрением данных. Наиболее распространены дискретное косинусное преобразование (*DCT*), отвечающее за преобразования дискретных данных в комбинации косинусных волн, вейвлет-преобразование (*DWT*), которое представляет собой математическое преобразование, позволяющее перевести сигнал из временного представления в частотно-временное и дискретное преобразование Фурье (*DFT*), которое переводит последовательность временных отсчётов сигнала в последовательность спектральных отсчётов.

После преобразования изображение представляется как совокупность частотных коэффициентов, отражающих энергию различных составляющих сигнала (яркость, детали, шум). Скрытая информация встраивается в эти коэффициенты — как правило, в среднечастотные области, где

изменения менее заметны для восприятия, но не отбрасываются при сжатии, как это происходит с высокочастотными компонентами.

Классы частотных методов можно разделить на внедрение информации в отдельные коэффициенты *DCT* или *DWT*, изменение относительных значений пары коэффициентов и манипулирование фазой или амплитудой преобразованного сигнала.

Преимущества данных методов заключаются в устойчивости к компрессии, трудности визуального и статистического обнаружения и поддержки форматов с потерями, а недостатки заключаются в сложности реализации, меньшей вместимости по сравнению с пространственными методами и в необходимости точного знания структуры преобразования при извлечении данных.

Таким образом, частотные методы считаются более надёжными и широко применяются в практической цифровой стеганографии, особенно для изображений, распространяемых в интернете, где доминирует формат *JPEG*.

1.5. Метод Коха–Жао в контексте формата *JPEG*

Одним из наиболее надёжных и устойчивых методов скрытия информации в цифровых изображениях является метод Коха–Жао, который реализуется на основе дискретного косинусного преобразования (*DCT*) — ключевого этапа обработки изображений в формате *JPEG*. Для полного понимания механизма внедрения информации этим методом, необходимо рассмотреть как работает *JPEG*-сжатие, каким образом применяется *DCT*, и где именно в этом процессе осуществляется внедрение скрытых данных [5].

1. Формат JPEG и структура изображения.

Формат *JPEG* (*Joint Photographic Experts Group*) использует сжатие с потерями, основанное на анализе восприятия человеком визуальной информации. Основные этапы кодирования изображения в этом формате следующие [6]:

- **Преобразование цветового пространства** из *RGB* в *YCbCr* (*Y* — яркость изображения, *Cb* — разница между синим компонентом и яркостью. *Cr* — разница между красным компонентом и яркостью), при этом компонент *Y* (яркость) отделяется от *Cb* и *Cr* (цветовых компонент), так как глаз более чувствителен к яркости.
- **Деление изображения на блоки 8×8 пикселей**, обрабатываемые отдельно. Это позволяет локализовать обработку и упростить сжатие.
- **Применение двумерного дискретного косинусного преобразования (*2D-DCT*)** к каждому блоку для перевода пиксельных значений из пространственной области в частотную.
- **Квантование** *DCT*-коэффициентов с помощью матрицы квантования — значения делятся и округляются, отбрасываются незначимые частоты.
- **Кодирование** для дополнительного уменьшения размера файла.

Наиболее важной частью этого процесса для задач стеганографии является 3-й этап — применение *DCT*, поскольку именно в его результате формируется набор коэффициентов, пригодный для внедрения информации.

2. Понятие дискретного косинусного преобразования.

Дискретное косинусное преобразование *(DCT)* — это математическое преобразование, которое выражает конечную последовательность данных через сумму косинусных функций с различными частотами. Также широко используется в стандартах сжатия, таких как *JPEG* (для изображений), *MPEG* (для видео), а также в аудиокодеках.

Основное назначение *DCT* заключается в преобразовании данных из пространственной области в частотную область, что позволяет:

- анализировать частотные характеристики сигналов и изображений;
- выполнять сжатие данных путем удаления менее значимых высокочастотных компонент;
- фильтровать шумы и улучшать качество сигналов;
- оптимизировать передачу и хранение мультимедийной информации.

Двумерное дискретное косинусное преобразование **(***2D DCT*) представляет собой расширение одномерного *DCT* для обработки двумерных данных, таких как цифровые изображения. В отличие от одномерного *DCT*, которое работает с последовательностями данных (например, аудиосигналами), *2D DCT* обрабатывает матрицы данных.

Двумерность означает, что преобразование применяется к данным, имеющим два пространственных измерения:

- первое измерение — строки изображения (горизонтальное направление);
- второе измерение — столбцы изображения (вертикальное направление).

Это позволяет анализировать частотные характеристики изображения как по горизонтали, так и по вертикали, что особенно важно для эффективного сжатия изображений, выделения текстур и границ объектов, фильтрации двумерных шумов, анализа пространственных частот в изображениях.

Математическое описание преобразования *2D DCT:*

Вычисление *2D DCT/IDCT* для матрицы пикселей размером $N \times N$ может быть определено как [7]:

$$y(u, v)= \frac{2c(u)c(v)}{N}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} x(i, j)\cos\frac{(2i+1)u\pi}{2N}\cos\frac{(2j+1)v\pi}{2N}, \quad (1)$$

Обратное преобразование (*2D IDCT*)8

$$x(i, j) = \frac{2}{N}\sum_{u=0}^{N-1}\sum_{v=0}^{N-1} c(u)c(v)y(u, v)\cos\frac{(2i+1)u\pi}{2N}\cos\frac{(2j+1)v\pi}{2N}, \quad (2)$$

Коэффициенты нормализации:

$$c(u), c(v)= \begin{cases} \sqrt{\frac{1}{2}}\ u, \text{ при } v=0 \\ 1u, \text{ при } v=1, 2, \ldots, N-1\ . \end{cases} \quad (3)$$

Двумерное *DCT/IDCT* разлагается на две последовательности, соответствующие строкам и столбцам, которые преобразуются дважды вдоль направлений строк и столбцов соответственно. Вычисление одномерного *DCT/IDCT* задается как:

Прямое одномерное *DCT*:

$$y(u) = \sqrt{\frac{2}{N}}\ c(u)\sum_{i=0}^{N-1} x(i)\cos\frac{(2i+1)u\pi}{2N}, \quad (4)$$

Обратное одномерное *IDCT*:

$$x(i)= \sqrt{\frac{2}{N}}\sum_{i=0}^{N-1} c(u)y(u)x(i)\cos\frac{(2i+1)u\pi}{2N}, \quad (5)$$

где:

- ***x(i, j)*** — исходное изображение (пиксельные значения);
- ***y(u, v)*** — размер каждой частотной компоненты после преобразования;
- ***N*** — размер матрицы ($N \times N$);
- ***i, j*** — пространственные координаты ($0 \le i, j \le N-1$);
- ***u, v*** — частотные координаты ($0 \le u, v \le N-1$);
- ***c(u), c(v)*** — коэффициенты нормализации.

Принцип работы заключается в том, что двумерное *DCT/IDCT* разлагается на две последовательности, соответствующие строкам и столбцам, которые преобразуются дважды: сначала вдоль направления строк, затем вдоль направления столбцов.

Внедрение данных в *DCT*-область методом Коха-Жао:

Стеганографический метод Коха-Жао представляет собой алгоритм скрытой передачи информации, основанный на применении двумерного дискретного косинусного преобразования (*2D DCT*) к цифровым изображениям. Данный метод позволяет встраивать секретную информацию в изображение-контейнер таким образом, чтобы визуальные изменения были минимальными и незаметными для человеческого глаза.

Основная идея метода заключается в модификации коэффициентов *DCT* в среднечастотной области спектра изображения. Выбор именно среднечастотных коэффициентов обусловлен необходимостью достижения компромисса между устойчивостью встроенной информации и незаметностью внесенных изменений.

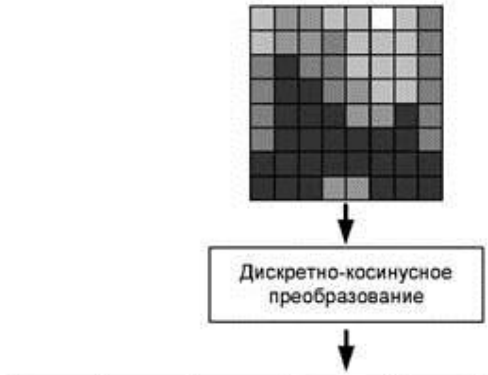
Процесс встраивания секретного сообщения в изображение-контейнер осуществляется в несколько этапов:

1. *Подготовка изображения.*

Исходное изображение разделяется на неперекрывающиеся блоки размером 8×8 пикселей. Данный размер блока является стандартным для алгоритмов, использующих *DCT*, и обеспечивает оптимальное соотношение между вычислительной эффективностью и качеством обработки.

2. *Применение дискретного косинусного преобразования.*

К каждому блоку изображения применяется двумерное *DCT*, в результате чего получаются матрицы коэффициентов *Di* (где i = 1, ..., *N*, *N* – общее количество блоков) размером 8×8. Коэффициенты *DCT* представляют частотные характеристики соответствующего блока изображения (см. рис. 2).



15,6	11,8	7,6	0	4	2,6	0	0
20,6	8,1	5,4	3,2	0	0	0	0
11,6	12,4	0	0	0	0	0	0
30,8	10,8	4,2	0	0	0	2,3	0
6,8	0,5	0	0	0	0,3	0	0
0	0	0	0,7	0	1,3	0	0,5
1,2	0	0	0,6	1,2	0	0	0
0	0	0,5	0	0	0	0	0

Рисунок 1. Преобразование DCT

3. *Выбор блоков для встраивания.*

Определяется последовательность блоков, которые будут использованы для встраивания секретной информации. В каждый выбранный блок встраивается один бит секретного сообщения, что обеспечивает контролируемую скорость встраивания.

4. *Выбор коэффициентов для модификации.*

В каждом блоке выбираются пары коэффициентов *DCT*, расположенные в среднечастотной области и симметричные относительно главной диагонали матрицы. Типичными парами являются *Di*[3, 4] и *Di*[4, 3], *Di*[3, 5] и *Di*[5, 3], *Di*[4, 5] и *Di*[5, 4].

Выбор среднечастотных коэффициентов критически важен, поскольку модификация низкочастотных коэффициентов приводит к заметным изменениям общей яркости и контрастности изображения, а изменение высокочастотных коэффициентов может вызвать появление артефактов и шумов.

5. *Встраивание информационных битов.*

Встраивание осуществляется путем модификации разности модулей выбранных пар коэффициентов:

для передачи бита «0»: разность модулей пары коэффициентов должна превышать положительное пороговое значение *M0*;

для передачи бита «1»: разность модулей должна быть меньше отрицательного порогового значения *-M0*.

Практическая реализация данного условия осуществляется следующим образом: при встраивании «0»: увеличивается модуль первого коэффициента и уменьшается модуль второго, а при встраивании «1»: уменьшается модуль первого коэффициента и увеличивается модуль второго

6. *Обработка всех блоков.*

Описанная процедура модификации коэффициентов выполняется последовательно для всех выбранных блоков изображения.

7. *Восстановление изображения.*

К каждому модифицированному блоку применяется обратное *DCT*, в результате чего получается стеганографическое изображение с встроенной секретной информацией.

Алгоритм извлечения скрытой информации из стеганографического изображения включает следующие этапы:

1. *Предварительная обработка.*

Первые этап алгоритма извлечения полностью идентичен соответствующим этапам алгоритма встраивания: разбиение на блоки, применение *DCT*, выбор блоков и коэффициентов.

2. *Вычисление разностей.*

Для каждой пары коэффициентов, использованной при встраивании, вычисляется разность модулей этих коэффициентов.

3. *Определение встроенных битов.*

На основе знака и абсолютного значения вычисленной разности определяется значение встроенного бита: если разность больше *M0*, то извлекается бит «0», а если разность меньше *-M0*, то извлекается бит «1».

4. *Формирование секретного сообщения.*

Биты, последовательно извлеченные из всех обработанных блоков, объединяются для восстановления исходного секретного сообщения.

Успешность атаки на стеганографический метод Коха-Жао зависит от возможности определения нескольких ключевых параметров:

- идентификация блоков с встроенной информацией – злоумышленник должен установить, какие именно блоки изображения содержат скрытые данные;
- определение порогового значения *M0* – критический параметр, влияющий на правильность извлечения информации;
- установление индексов модифицированных коэффициентов *DCT* – необходимо знать точные позиции коэффициентов, использованных для встраивания.

Сложность определения этих параметров без знания ключевой информации обеспечивает определенный уровень безопасности метода, хотя стойкость алгоритма может быть повышена за счет использования дополнительных мер защиты, таких как криптографическое шифрование встраиваемых данных или применение псевдослучайных последовательностей для выбора блоков и коэффициентов.

2. Разработка комплекса по детектированию

2.1. Проектирование системы

Проектирование началось с разработки архитектурных диаграмм, которые стали основой для всей последующей работы:

1. *Mind Map системы (см. рис. 3).*

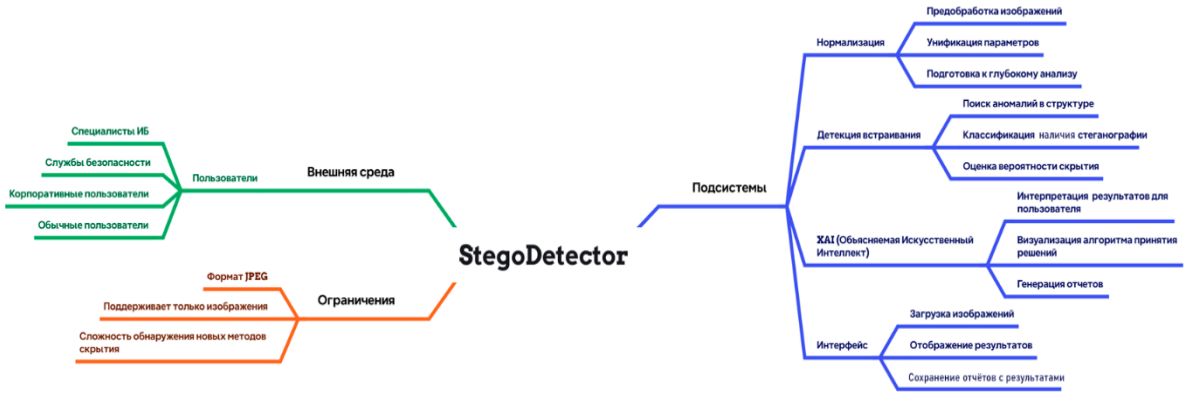


Рисунок 2. Mind map системы

Диаграмма представляет собой общее концептуальное представление подсистем, а также демонстрирует целевую аудиторию и ограничения системы. Среди пользователей выделены службы безопасности, корпоративные пользователи и обычные пользователи.

Система включает следующие взаимосвязанные подсистемы:

- Подсистема нормализации изображений:

Отвечает за приведение изображений к единому виду, включая стандартизацию цветовой схемы, размеров и удаление лишней сопутствующей информации, которая может повлиять на дальнейший анализ.

- Подсистема детектирования скрытых данных:

Осуществляет извлечение характерных признаков из изображений и их последующую обработку с помощью методов машинного обучения, направленных на выявление возможных скрытых изменений или аномалий.

- Подсистема интерпретации результатов:

Предназначена для повышения прозрачности работы моделей. Позволяет визуализировать, какие признаки повлияли на итоговые решения, и формировать понятные объяснения для пользователей.

- Пользовательский интерфейс:

Обеспечивает взаимодействие с системой, включая загрузку изображений, выбор режимов анализа, просмотр результатов и доступ к справочной информации, описывающей принципы функционирования и особенности используемых методов.

2. *BPMN-диаграмма бизнес-процесса:*

Описывает этапы взаимодействия пользователя с системой (см. рис. 4).

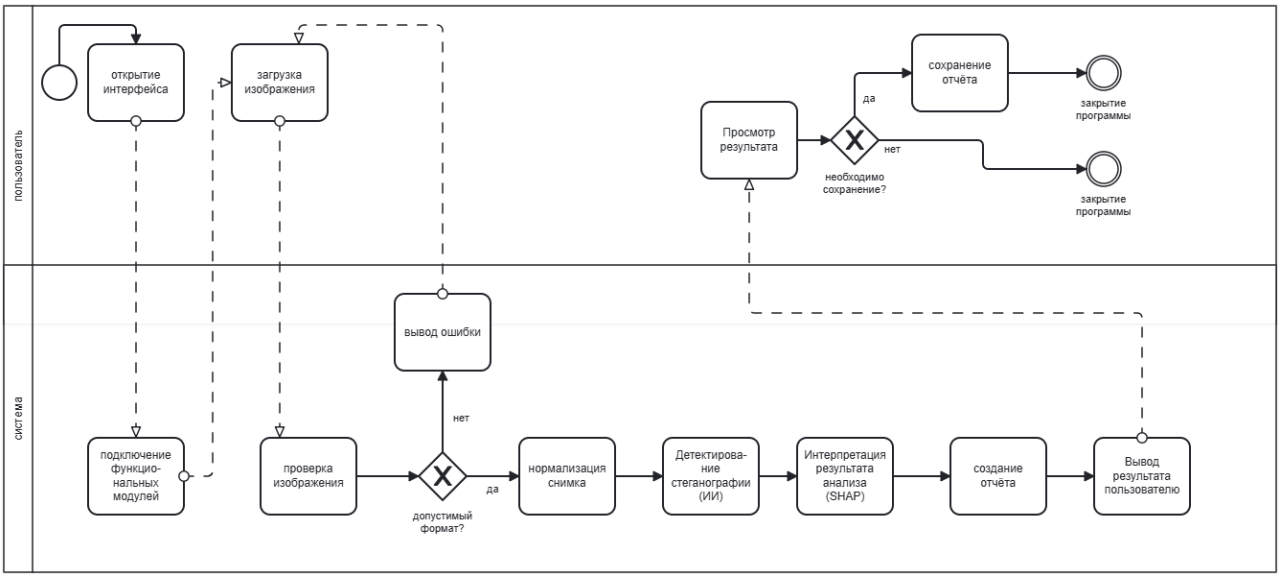


Рисунок 3. BPMN-диаграмма

2.2. Сбор датасета

Для создания датасета были выбраны изображения животных (кошек и собак), загруженные из открытого источника, так как предпочтение отдавалось реальным фотографиям с естественным освещением и разнообразными текстурами фона. Всего было собрано 3800 изображений в формате *JPG* и произведены следующие действия:

- Нормализация изображений – все изображения были приведены к унифицированному размеру, применено преобразование к оттенкам серого для упрощения частотного анализа, из изображений были удалены метаданные (*EXIF*-информация), чтобы предотвратить нежелательные вариации.

- Генерация пар «чистое» / «модифицированное» изображение – каждое исходное изображение использовалось дважды – в оригинальной форме (без модификаций) и с внедрённым сообщением методом Коха-Жао. В качестве сообщения использовалась заранее подготовленная текстовая строка, переведённая в бинарный массив.

- Формирование датасета – итоговый датасет содержал равное количество «чистых» и «модифицированных» изображений. Каждое изображение было снабжено соответствующей меткой класса («*without_embedding*» или «*with_koha*»).

2.3. Создание модуля детектирования скрытых данных

Метод встраивания:

В изображение внедрялось сообщение методом Коха-Жао, модифицируя среднечастотные *DCT*-коэффициенты с случайно генерируемым параметром *alpha* от 0.4 до 0.6 (сила внедрения).

Признаки на основе *DCT*-коэффициентов:

- среднее значение (*dct_mean*);
- стандартное отклонение (*dct_std*);
- коэффициент асимметрии (*dct_skewness*);
- эксцесс (*dct_kurtosis*).

Признаки на основе разностей между блоками:

- среднее значение горизонтальных разностей (*diff_horiz_mean*);
- стандартное отклонение горизонтальных разностей (*diff_horiz_std*);
- среднее значение вертикальных разностей (*diff_vert_mean*);

- стандартное отклонение вертикальных разностей (*diff_vert_std*).

Признаки на основе анализа шума:

- среднее значение шума (*noise_mean*);
- стандартное отклонение шума (*noise_std*).

Гистограмма распределения *DCT*-коэффициентов:

- 10 нормализованных бинов от -150 до 150.

Эти признаки позволяют обнаружить микроскопические изменения в частотной структуре изображения, вызванные внедрением скрытых сообщений.

Для решения задачи бинарной классификации («наличие стеганографического вмешательства» / «отсутствие вмешательства») была разработана и обучена полносвязная нейронная сеть (*Fully Connected Neural Network, FCNN*) [8].

Ключевые особенности архитектуры нейронной сети:

- количество входных признаков: 20 признаков на изображение;
- количество скрытых слоёв: 4 полносвязных слоя;
- функция активации: *ReLU* (*Rectified Linear Unit*) на скрытых слоях;
- функция активации на выходе: *Sigmoid* (для бинарной классификации);
- функция потерь: Binary Cross-Entropy Loss;
- оптимизатор: Adam (Adaptive Moment Estimation);

Структура модели:

- входной слой: 20 нейронов (по одному на каждый извлечённый признак);
- скрытый слой 1: 64 нейрона + активация *ReLU*;
- скрытый слой 2: 32 нейрона + активация *ReLU*;
- скрытый слой 3: 16 нейронов + активация *ReLU*;
- скрытый слой 4: 8 нейронов + активация *ReLU*;
- выходной слой: 1 нейрон с активацией *Sigmoid*.

Первичное обучение и тестирование модели детектирования:

После обучения на 30 эпохах модель показала точность (*Accuracy*) на уровне 78,2%.

На диаграмме ниже можно увидеть функцию потерь и точности при обучении (см. рис. 5).

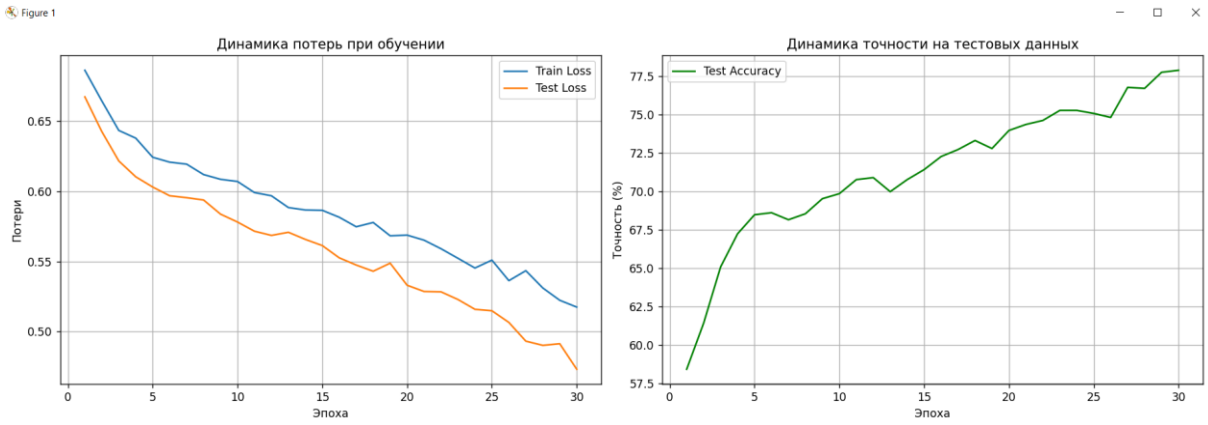


Рисунок 5. Диаграмма динамик потерь и точности

Анализ ошибок классификации:

- Ошибка I рода (*False Positive Rate*) – 21,15%: чистые изображения ошибочно определены как содержащие стеганографию.
- Ошибка II рода (*False Negative Rate*) – 23,11%: случаи скрытой информации остались нераспознанными.

Дополнительные метрики:

- *Precision* (Точность положительного класса) – 78,43%;
- *Recall* (Полнота) – 76,89%;
- *F1*-мера – 77,65%, отражает сбалансированность между точностью и полнотой;
- *AUC-ROC* – 86,04%, свидетельствует о высоком качестве бинарной классификации.

2.4. Интеграция технологий объяснительного ИИ (*XAI*)

Интеграция библиотеки *SHAP* [9]:

- для каждого анализа формируются объяснения — какие признаки повлияли на решение;
- вывод графиков важности признаков при каждом выводе.

Использованная технология позволяет отобразить пользователю, как именно модель пришла к решению о наличии (или отсутствии) стеганографического вмешательства).

Следующая страница разработанного оконного приложения предназначена для отображения влияния различных признаков на обнаружение стеганографии с использованием технологии *SHAP* (*SHapley Additive exPlanations*) (см. рис. 6). На странице строится диаграмма, где:

- красные столбцы указывают на признаки, увеличивающие вероятность обнаружения стеганографии;
- синие столбцы показывают признаки, снижающие вероятность обнаружения;
- длина столбцов соответствует степени влияния признака.

Рисунок 6. Страница со значениями SHAP

1. *Страница DCT блоков.*

На этой странице отображается тепловая карта энергии *DCT* блоков, которая визуализирует распределение энергии в дискретном косинусном преобразовании изображения(см. рис. 7). Данная визуализация позволяет:

- определить области изображения с необычным распределением энергии *DCT*;
- выявить аномалии, характерные для стеганографических вставок;
- проанализировать энергетический "отпечаток" изображения.

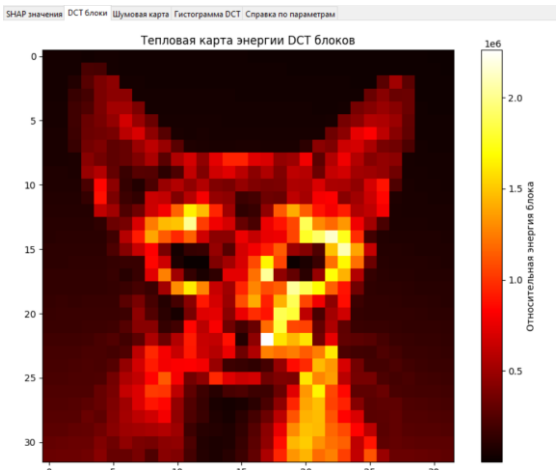


Рисунок 7. Страница с тепловой картой энергии DCT блоков

2. *Страница шумовой карты.*

Страница содержит две основные компоненты:

- карту шумов (*DCT*-анализ) - визуализацию распределения шумов в изображении;
- гистограмму распределения шума с пороговым значением.

Данная страница позволяет выявить аномальные шумовые характеристики, которые могут свидетельствовать о встраивании скрытой информации (см. рис. 8).

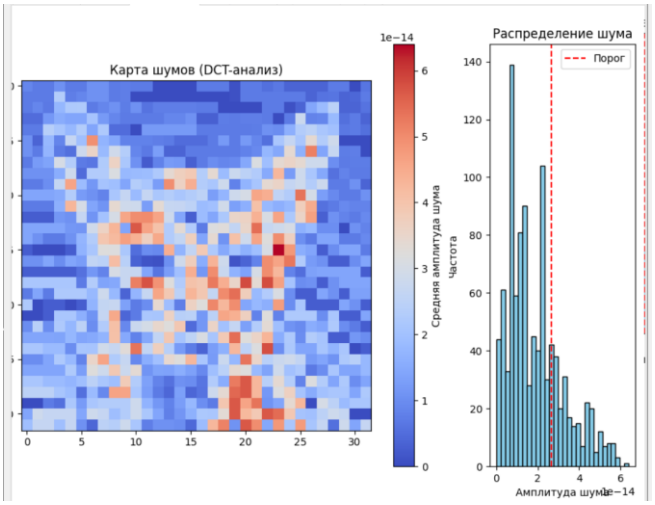


Рисунок 8. Страница с картой шумов

Заключение

В ходе выполнения работы была разработана и реализована система детектирования стеганографических сообщений в графических изображениях с применением технологий объяснимого искусственного интеллекта.

Проанализированы существующие подходы к стегоанализу, включая как классические статистические методы, так и современные алгоритмы машинного обучения. Отдельное внимание уделено возможностям применения методов объяснимого ИИ для повышения прозрачности и доверия к работе моделей. В рамках исследования был собран специализированный датасет, содержащий как оригинальные изображения, так и изображения с внедрённой скрытой информацией. На его основе разработана архитектура нейронной сети и построен классификатор, демонстрирующий высокую точность в определении наличия стеганографических сообщений. Интеграция модулей объяснимого ИИ, в частности библиотеки *SHAP*, обеспечила

интерпретируемость результатов и визуализацию вклада отдельных признаков в итоговую классификацию.

Результаты работы подтверждают эффективность предложенного подхода к детектированию стеганографических сообщений в *JPEG*-изображениях. Разработанный программный комплекс может быть востребован в задачах информационной безопасности, экспертных исследованиях цифрового контента и при проведении судебных экспертиз в области кибербезопасности [10].

Список источников

1. Гребенников В. Стеганография. История тайнописи. – Москва: ЛитРес: Самиздат, 2024. – 142 с.
2. Грибунин В. Г., Оков И. Н., Туринцев И. В. Цифровая стеганография. – Москва: СОЛОН-ПРЕСС, 2009. – 264 с.
3. Абазина Е. С., Ерунов А. А. Цифровая стеганография: состояние и перспективы // Системы управления, связи и безопасности. – 2016. – № 2. – С. 182-201.
4. Половинченко М. И., Елисеев В. С. Звуковые данные и функции преобразования Фурье, БПФ и спектрограмм для системы распознавания речи // Journal of Advanced Research in Technical Science. – 2021. – №. 25. – С. 74-81.
5. Вильховский Д. Э. Стеганографический анализ изображений на предмет обнаружения вставок, выполненных методом Коха-Жао // Математическое и компьютерное моделирование. Сборник материалов IX 126 Международной научной конференции, посвященной 85-летию профессора В.И. Потапова.– Омск, 2021.– С. 319-321.
6. Iqbal Y., Kwon O. J. Improved JPEG coding by filtering 8× 8 DCT blocks // Journal of Imaging. – 2021. – Т. 7. – №. 7. – С. 117.
7. Zhou Z., Pan Z. Effective hardware accelerator for 2d dct/idct using improved loeffler architecture // IEEE Access. – 2022. – Т. 10. – С. 11011-11020.
8. Han Y., Hong B. W. Deep learning based on fourier convolutional neural network incorporating random kernels //Electronics. – 2021. – Т. 10. – №. 16. – С. 2004.
9. Аверкин А. Н., Ярушев С. А. Объяснительный искусственный интеллект в моделях поддержки принятия решений для Здравоохранения 5.0 // Компьютерные инструменты в образовании. – 2023. – №. 2. – С. 41-61.
10. Станчук П. Н. Применение стеганографии в компьютерных атаках //Вестник науки и образования. – 2023. – №. 6 (137). – С. 29-33.