Detecção de esteganografia em imagens utilizando aprendizado de máquina

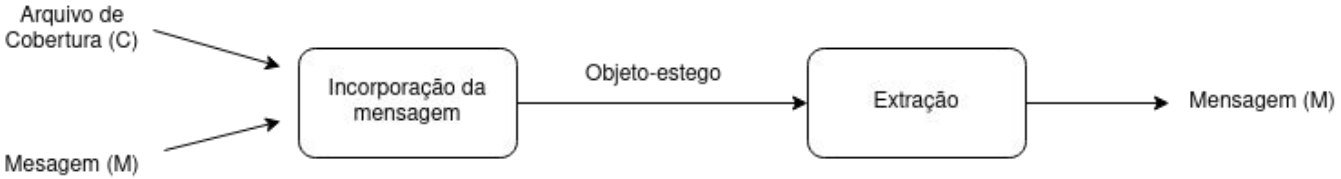
Aluno: Matheus Esquinelato Polachini
Orientador: Prof. Dr. Kelton Augusto Pontara da Costa

Trabalho de Conclusão de Curso
Bacharelado em Ciência da Computação
Faculdade de Ciências - UNESP - Bauru

Conteúdo

Introdução

- Esteganografia se refere ao processo de esconder uma mensagem em um meio de forma a ser imperceptível para quem não tenha conhecimento prévio da existência da mensagem
- Esteganografia digital se refere ao processo de esconder a mensagem em arquivos de mídia digital, sendo mais comum a utilização de imagens



Fonte: Adaptado de Silva, Carvalho e Martins (2020)

Introdução

- Esteganálise é o nome dado ao processo de detecção de esteganografia.
- Uma de suas vertentes, a análise estatística, se baseia no fato de que o processo de esteganografia causa alterações em propriedades estatísticas da imagem que podem ser detectadas
- Neste trabalho foi proposta a utilização de técnicas de aprendizado de máquina para análise de propriedades estatísticas da imagem de forma a detectar a presença de uma mensagem escondida através de esteganografia

Introdução - Problema e Justificativa

- Com o aumento na utilização de métodos de esteganografia e com o contínuo desenvolvimento de novos métodos, há a necessidade de novas maneiras de detecção de esteganografia
- Existem várias ferramentas acessíveis publicamente para aplicar esteganografia em imagens. Apesar dessa facilidade de uso ser benéfica para usos legais, organizações criminosas já utilizaram esteganografia para se comunicar sem levantar suspeitas

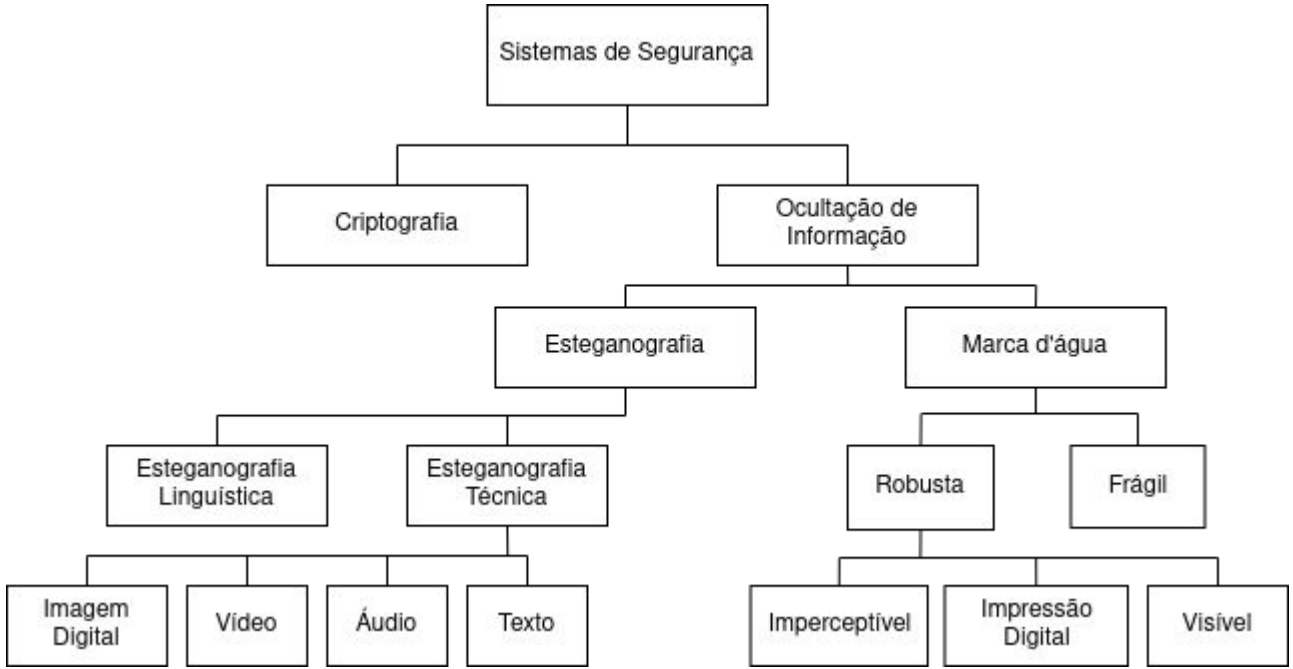
Introdução - Objetivos

- Estudar conceitos e técnicas de aprendizado de máquina e de esteganografia
- Reunir um conjunto de dados adequado para o treinamento do modelo de classificação
- Definir a arquitetura e a técnica utilizada no classificador
- Implementar o classificador e treiná-lo utilizando os dados coletados
- Avaliar o desempenho e a taxa de acerto do classificador

Fundamentação - Técnicas de Segurança

- Técnicas desenvolvidas com o objetivo de estabelecer uma comunicação secreta entre duas partes existem desde a época da Grécia Antiga
- Atualmente, com a predominância da comunicação por meios digitais, novas técnicas de segurança de informação foram desenvolvidas para atingir variados objetivos
- Essas técnicas podem ser classificadas em criptografia, esteganografia e marca d'água.

Fundamentação - Técnicas de Segurança



Fonte: Adaptado de Kadhim et al. (2019)

Fundamentação - Técnicas de Segurança

- Técnicas de criptografia têm como objetivo transformar a informação em uma forma que é compreensível apenas para o emissor e o receptor
- Um sistema de criptografia é ineficaz quando um terceiro tem acesso à mensagem original, enquanto um sistema de esteganografia é ineficaz quando um terceiro detecta a presença da mensagem
- Técnicas de marca d'água digital são usadas para identificar o criador, dono, distribuidor ou consumidor autorizado de um documento, tendo como prioridade a robustez

Fundamentação - Técnicas de Segurança

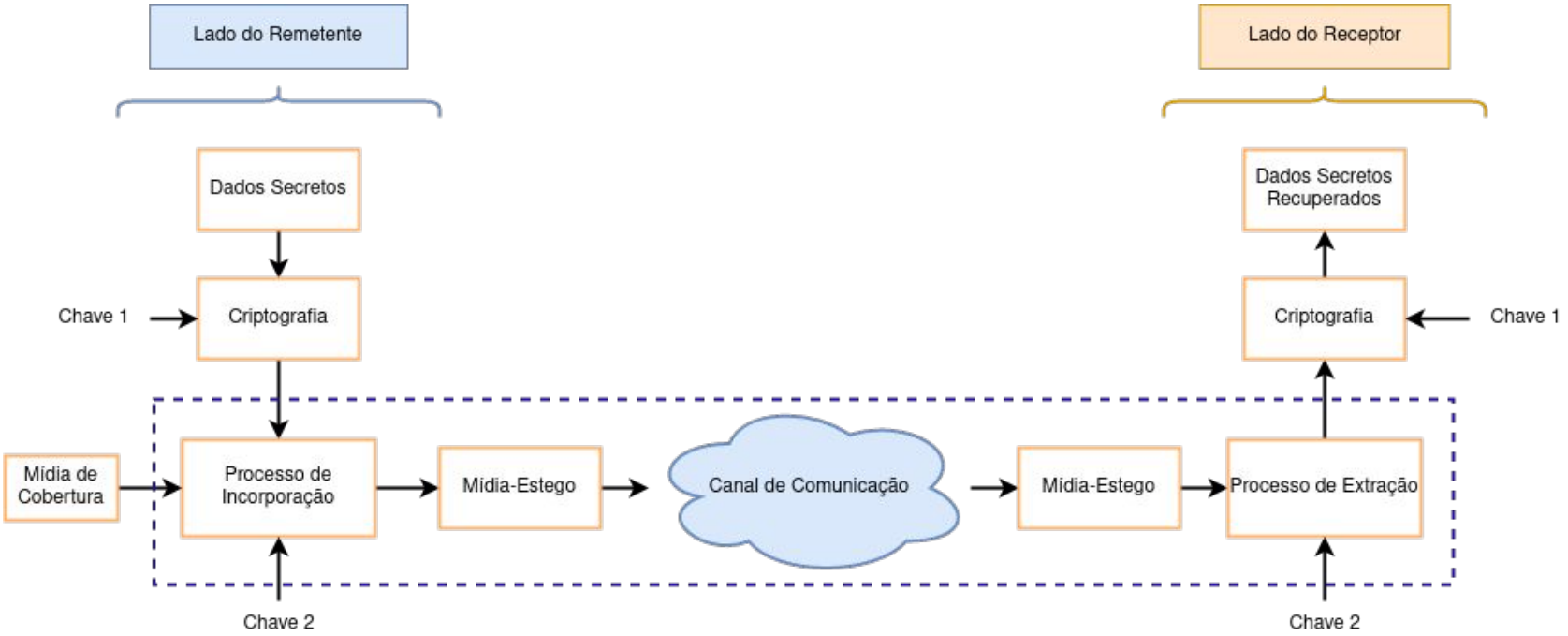
Característica	Esteganografia	Marca d'água	Criptografia
Objetivo	Evitar que o dado confidencial seja detectado	Preservar a autenticidade do arquivo de mídia	Ofuscar a forma ou conteúdo da mensagem
Escolha do arquivo	Livre	Restrita	-
Desafios	Imperceptibilidade	Robustez	Robustez
Chave	Opcional	Opcional	Obrigatória
Visibilidade	Não visível	Visível em alguns casos	Sempre visível
Inválido se	Detectado	Removido ou substituído	Decifrado
Ataques	Esteganálise	Qualquer processamento de imagem	Criptoanálise

Fonte: Adaptado de (KADHIM et al., 2019)

Fundamentação - Esteganografia Digital

- Arquivos de imagem são os mais utilizados por conterem um grande número de bits redundantes e por serem facilmente compartilháveis em diversos serviços através da internet
- É possível combinar a criptografia e a esteganografia ao esconder a mensagem criptografada na imagem
- Existem três propriedades essenciais de um sistema de esteganografia: segurança, capacidade e robustez

Fundamentação - Esteganografia Digital



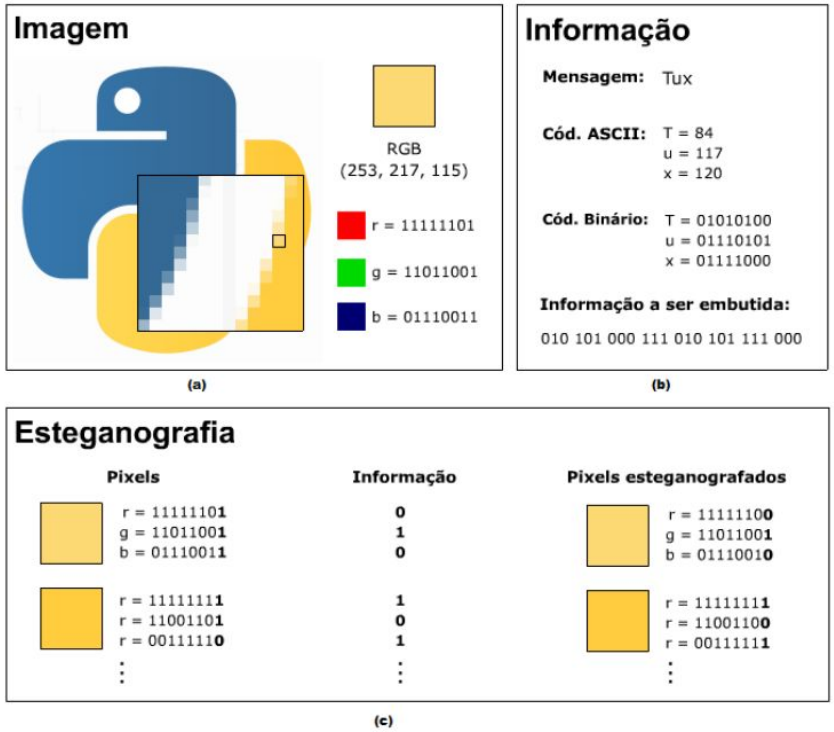
Fundamentação - Esteganografia Digital

- Formas de classificação das técnicas de esteganografia:
  - quanto à dimensão da imagem de cobertura
  - quanto à natureza de recuperação
  - quanto ao domínio de incorporação
  - esteganografia adaptativa
- A classificação mais frequente na literatura é baseada no domínio de incorporação, subdividindo-se em técnicas que utilizam o domínio espacial e técnicas que utilizam o domínio de frequência
- Técnicas baseadas no domínio espacial funcionam apenas em imagens que utilizam um formato de compressão sem perda, sendo o mais comum o PNG

Fundamentação - Substituição LSB

- Consiste na utilização dos bits menos significativos dos valores de pixel da imagem para armazenar a mensagem secreta
- Essa técnica se baseia no fato de que a alteração dos bits menos significativos dos pixels é imperceptível pela visão humana
- Existem variações dessa técnica, por exemplo definindo a posição de início da mensagem ou incorporando os bits de maneira não sequencial
- Um estudo de 2011 demonstrou que 70% dos softwares de esteganografia utilizam o algoritmo LSB ou uma de suas variações

Fundamentação - Substituição LSB



Fonte: Sgursky (2015)

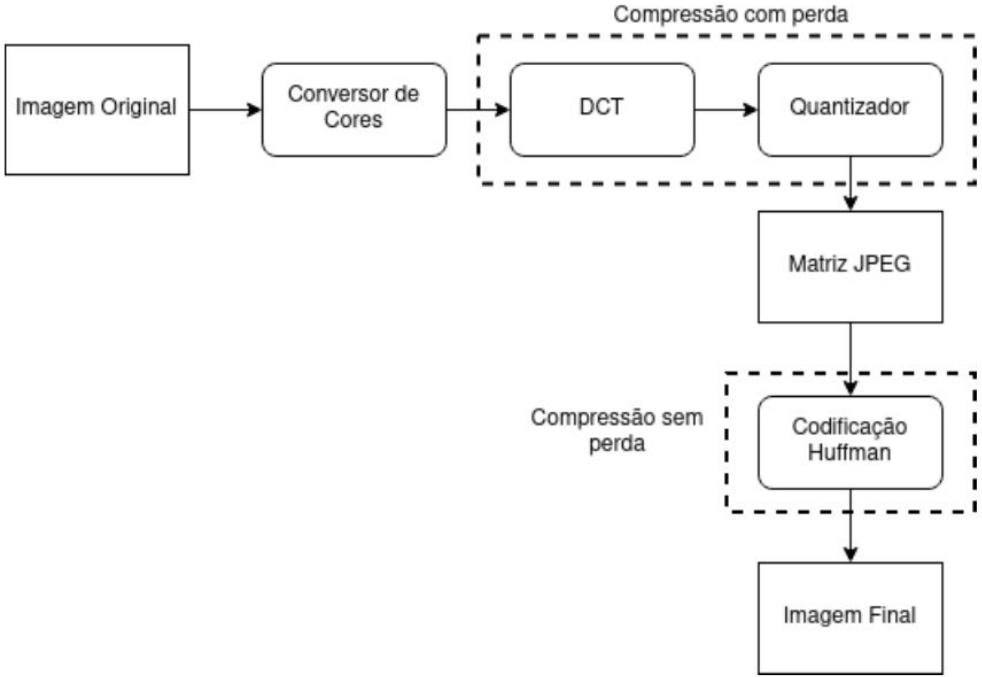
Fundamentação - Diferença de Valor de Pixel

- Consiste em esconder a mensagem através da comparação das diferenças dos valores de dois pixels sucessivos.
- A imagem é dividida em uma série de grupos com dois pixels adjacentes entre eles. A diferença dos valores dos pixels adjacentes de cada grupo é calculada e substituída por uma parte da mensagem secreta.
- Utilizando essa técnica, é possível alterar a imagem apenas em uma certa faixa de valores de diferença de pixel

Fundamentação - JSteg

- Usado em imagens JPEG e baseado na Transformada Discreta de Cosseno
- Durante o processo de compressão JPEG, cada componente de cor da imagem é dividido em blocos de tamanho 8x8. Cada bloco é mapeado através de uma Transformada de Cosseno produzindo um bloco 8x8 de coeficientes DCT.
- O algoritmo JSteg consiste em realizar a substituição dos bits menos significativos dos coeficientes DCT.

Fundamentação - JSteg



Fundamentação - Métricas de Qualidade da Imagem

- Conjunto de métricas que possuem o objetivo de comparar uma imagem com outra imagem distorcida ou adulterada e avaliar a diferença entre elas
- Foi sugerido por Sgursky (2015) a utilização dessas métricas como características para detecção de esteganografia. Nesse caso, como existe acesso apenas a uma imagem, a comparação é feita entre essa imagem e a mesma imagem convertida para outro formato (por exemplo do PNG para o JPEG)
- Outra forma de utilização dessas métricas para detecção de esteganografia é realizando a comparação entre uma imagem e a mesma imagem com um filtro gaussiano aplicado, método sugerido por Schaathun (2012).

Fundamentação - Métricas de Qualidade da Imagem

- Distância Média

$$\sum_{j=1}^{M}\sum_{k=1}^{N}(F(j,k)-G(j,k))/MN$$

- Distância Euclidiana

$$\frac{1}{MN}(\sum_{j=1}^{M}\sum_{k=1}^{N}(F(j,k)-G(j,k))^2)^{1/2}$$

- Conteúdo Estrutural

$$\sum_{j=1}^{M}\sum_{k=1}^{N}F(j,k)^2/\sum_{j=1}^{M}\sum_{k=1}^{N}G(j,k)^2$$

- Fidelidade da Imagem

$$1-(\sum_{j=1}^{M}\sum_{k=1}^{N}(F(j,k)-G(j,k))^2/\sum_{j=1}^{M}\sum_{k=1}^{N}G(j,k)^2)$$

Fundamentação - Métricas de Qualidade da Imagem

- Correlação Cruzada Normalizada

$$\sum_{j=1}^{M}\sum_{k=1}^{N} F(j,k)G(j,k) / \sum_{j=1}^{M}\sum_{k=1}^{N} F(j,k)^2$$

- Erro Médio Quadrático Normal

$$\sum_{j=1}^{M}\sum_{k=1}^{N} (F(j,k) - G(j,k))^2 / \sum_{j=1}^{M}\sum_{k=1}^{N} F(j,k)^2$$

- Sinal de Pico-Ruído

$$20 \times log_{10}\{255/\{\sum_{j=1}^{M}\sum_{k=1}^{N}[F(j,k) - G(j,k)]^2\}^{1/2}\}$$

Fundamentação - Métricas de Qualidade da Imagem

- Erro Médio Quadrático Mínimo

$$\sum_{j=1}^{M-1}\sum_{k=2}^{N-1}(F(j,k)-G(j,k))^2 / \sum_{j=1}^{M-1}\sum_{k=2}^{N-1}O(F(j,k))^2$$

$$O(F(j,k)) = F(j+1,k) + G(j-1,k) + F(j,k+1) + F(j,k-1) - 4F(j,k)$$

- Pico do Erro Médio Quadrático

$$\frac{1}{MN}\sum_{j=1}^{M}\sum_{k=1}^{N}[F(j,k)-G(j,k)]^2 / \{max_{j,k}[F(j,k)]\}^2$$

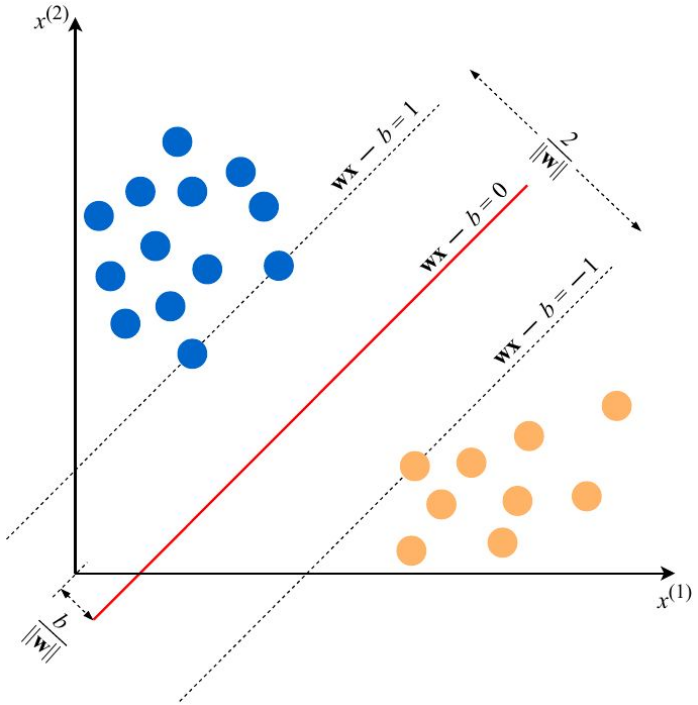
Fundamentação - Aprendizado de Máquina

- Consiste em reunir um conjunto de dados sobre determinado fenômeno e construir um modelo estatístico baseado nesse conjunto de dados
- Cada elemento é identificado por um vetor de características e pode ou não ter um rótulo associado
- O aprendizado pode ser de quatro tipos:
  - Aprendizado supervisionado
  - Aprendizado não supervisionado
  - Aprendizado semi supervisionado
  - Aprendizado por reforço:

Fundamentação - Aprendizado de Máquina

- Métricas de avaliação:
  - Matriz de confusão
  - Acurácia (taxa de elementos classificado corretamente em relação ao total)
  - Precisão (taxa de predições positivas corretas em relação ao total de predições positivas)
  - Recall (taxa de predições positivas corretas em comparação com o total de elementos positivos)

Fundamentação - Aprendizado de Máquina - SVM



Fonte: Burkov (2019)

Desenvolvimento

- Técnicas de esteganografia selecionadas: LSB, Diferença de Valor de Pixel e JSteg
- Foi utilizada a abordagem de construção de um classificador binário para cada técnica
- A técnica de aprendizado de máquina escolhida foi o SVM devido a ser a técnica mais usada em trabalhos relacionados

Desenvolvimento

- Foi utilizada uma base de imagens fornecida por Liu, Cooper e Zhou (2013) que contém 5150 imagens coloridas não compactadas de tamanho 256x256 pixels no formato BMP

Fonte: elaborado pelo autor

Desenvolvimento

- Ferramentas utilizadas:
  - Linguagens de programação Python e Go
  - Visual Studio Code
  - Google Colaboratory
  - Bibliotecas Python:
    - Numpy
    - Pandas
    - Pillow
    - OpenCV
    - Scikit-learn
    - Matplotlib

Desenvolvimento

- Conversão da base de imagens para os formatos PNG e JPEG
- Para cada técnica, foi realizado o processo de esteganografia em 50% das imagens, resultando em um conjunto com 2575 imagens originais e 2575 imagens com esteganografia
- A técnica LSB foi subdividida em 5 técnicas, sendo cada uma utilizando 10%, 25%, 50%, 75% e 100% da capacidade total da imagem

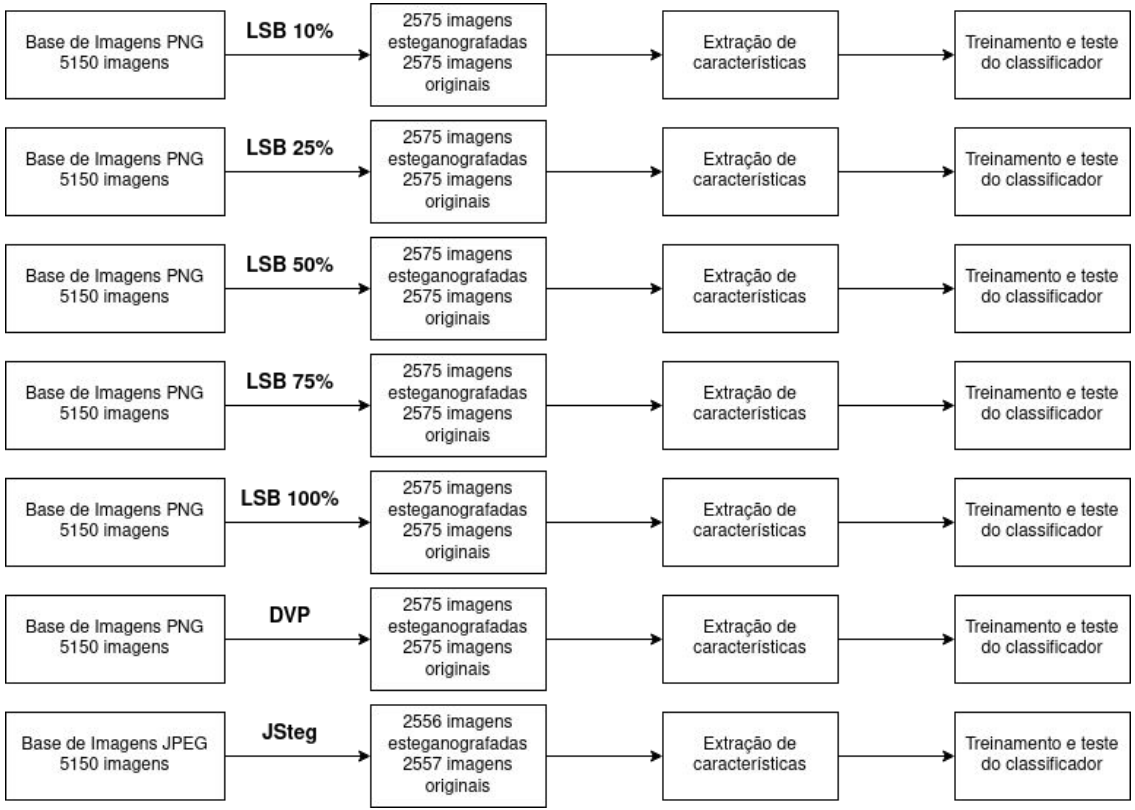
Desenvolvimento

- No caso da técnica Diferença de Valor de Pixel, foi utilizado o valor fixo de 1220 caracteres por imagem. Para o JSteg, foram utilizados 500 caracteres para imagens maiores que 10 KB, 100 caracteres para arquivos de tamanho entre 4 e 10 KB, 50 caracteres para arquivos entre 3 e 4 KB e 10 caracteres para os restantes
- No caso do JSteg, algumas imagens tiveram que ser descartadas por não possibilitarem a incorporação de uma mensagem de tamanho minimamente significativo, o que resultou na redução do conjunto de imagens de 5150 para 5113 imagens

Desenvolvimento

- Como características para o treinamento e teste do classificador, foram utilizadas as métricas de qualidade da imagem.
- No caso das técnicas LSB e Diferença de Valor de Pixel, o cálculo das métricas foi feito entre a imagem PNG e a mesma imagem convertida para o formato JPEG. No caso do JSteg, o cálculo foi feito entre a imagem JPEG e a mesma imagem com um filtro gaussiano aplicado.
- Para o processo de treinamento e teste do classificador, o conjunto de imagens de cada técnica foi dividido em 70% para treinamento e 30% para teste

Desenvolvimento



Fonte: elaborado pelo autor

Resultados

Técnica	Acurácia	Precisão	Recall
LSB 10%	53.33%	52.56%	55.55%
LSB 25%	57.22%	55.23%	64.54%
LSB 50%	67.77%	65.54%	73.76%
LSB 75%	72.56%	70.96%	76.29%
LSB 100%	79.16%	76.98%	82.53%
DVP	59.42%	60.50%	56.00%
JSteg	52.02%	54.66%	32.35%

Resultados

- Analisando os resultados dos classificadores LSB, é possível observar que todas as três métricas crescem à medida que o tamanho da mensagem aumenta, indicando que as métricas utilizadas como característica refletem a existência de esteganografia
- Os resultados dos classificadores de Diferença de Valor de Pixel e JSteg demonstraram uma eficácia muito menor em comparação com o menor caso do LSB. Acredita-se que isso ocorreu devido à menor capacidade dessas técnicas, o que levou à necessidade de utilização de uma mensagem menor em comparação com o LSB.

Resultados

- Em comparação com trabalhos similares, os resultados do LSB foram ligeiramente melhores que os obtidos por Schaathun (2012) ao analisar a acurácia de detecção LSB utilizando métricas de qualidade da imagem. O referido trabalho obteu acurácia de 62% para LSB utilizando 40% de capacidade e acurácia de 71.7% para LSB utilizando 100% de capacidade

Conclusão

- Neste trabalho foi avaliada a abordagem de uso de aprendizado de máquina para detecção de esteganografia em imagens
- Foi realizado um estudo bibliográfico com o objetivo de selecionar técnicas e métricas que possam ser utilizadas para treinar um classificador. Em seguida, os classificadores foram construídos utilizando a técnica SVM e extração de características por métricas de qualidade da imagem
- Os resultados sugerem que essa abordagem é eficaz nos casos em que a quantidade de informação incorporada na imagem é grande, porém perde sua eficiência nos casos em que a informação é mínima

Trabalhos futuros

- Análise da eficácia de detecção utilizando outras técnicas de esteganografia e de aprendizado de máquina
- Teste em outras bases de dados
- Estudo de métricas e características que possam detectar informações de tamanho pequeno incorporadas em uma imagem

Referências

- SILVA, W. G. da; CARVALHO, R. L. de; MARTINS, G. A. de S. Steganography genetic algorithm hyperparameter tuning through response surface methodology. *Academic Journal on Computing, Engineering and Applied Mathematics*, v. 1, n. 1, p. 13–17, 2020.
- SHIH, F. Y. *Digital watermarking and steganography: fundamentals and techniques*. [S.l.]:CRC press, 2017.
- SGURSKY, L. F. F. Análise e implementação de técnicas de esteganografia. 2015.
- KADHIM, I. J.; PREMARATNE, P.; VIAL, P. J.; HALLORAN, B. Comprehensive survey of image steganography: Techniques, evaluations, and trends in future research. *Neurocomputing*, Elsevier, v. 335, p. 299–326, 2019.
- SCHAATHUN, H. G. *Machine learning in image steganalysis*. [S.l.]: Wiley Online Library, 2012.
- LIU, Q.; COOPER, P. A.; ZHOU, B. An improved approach to detecting content-aware scaling-based tampering in jpeg images. In: IEEE. 2013 IEEE China Summit and International Conference on Signal and Information Processing. [S.l.], 2013. p. 432–436.
- BURKOV, A. *The hundred-page machine learning book*. [S.l.]: Andriy Burkov Quebec City, QC, Canada, 2019. v. 1.
- SHEISI, H.; MESGARIAN, J.; RAHMANI, M. Steganography: Dct coefficient replacement method and compare with jsteg algorithm. International Journal of Computer and Electrical Engineering, v. 4, n. 4, p. 458–462, 2012.