UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS - CAMPUS BAURU
DEPARTAMENTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MATHEUS ESQUINELATO POLACHINI

DETECÇÃO DE ESTEGANOGRAFIA EM IMAGENS UTILIZANDO APRENDIZADO DE MÁQUINA

BAURU
Agosto/2022

MATHEUS ESQUINELATO POLACHINI

DETECÇÃO DE ESTEGANOGRAFIA EM IMAGENS UTILIZANDO APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.
Orientador: Prof. Dr. Kelton Augusto Pontara da Costa

BAURU
Agosto/2022

Matheus Esquinelato Polachini Detecção de Esteganografia em Imagens Utilizando Aprendizado de Máquina/ Matheus Esquinelato Polachini. – Bauru, Agosto/2022- 33 p. : il. (algumas color.) ; 30 cm.
Orientador: Prof. Dr. Kelton Augusto Pontara da Costa
Trabalho de Conclusão de Curso – Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Bacharelado em Ciência da Computação, Agosto/2022.
1. Esteganografia 2. Aprendizado de Máquina 3. Sistemas de Segurança 4. Support Vector Machine

Matheus Esquinelato Polachini

Detecção de Esteganografia em Imagens Utilizando Aprendizado de Máquina

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.

Banca Examinadora

Prof. Dr. Kelton Augusto Pontara da Costa
Orientador
Departamento de Computação
Faculdade de Ciências
Universidade Estadual Paulista "Júlio de Mesquita Filho"

Profª. Drª. Simone das Graças Domingues Prado
Departamento de Computação
Faculdade de Ciências
Universidade Estadual Paulista "Júlio de Mesquita Filho"

Prof. Dr. Clayton Reginaldo Pereira
Departamento de Computação
Faculdade de Ciências
Universidade Estadual Paulista "Júlio de Mesquita Filho"

Bauru, 03 de Agosto de 2022.

Resumo

Esteganografia em imagens se refere ao processo de incorporação de uma mesagem secreta em um arquivo de imagem sem causar mudança visual perceptível a quem tenha acesso a essa imagem. Devido ao contínuo desenolvimento de novas técnicas de esteganografia, há a necessidade de desenvolvimento de novas formas de detecção dessas técnicas. Este trabalho buscou investigar a eficiência do uso de aprendizado de máquina na detecção das técnicas de esteganografia LSB, PVD e JSteg em imagens utilizando a técnica de aprendizado de máquina *Support Vector Machine* e características por métricas de qualidade da imagem.
Palavras-chave: Esteganografia, *SVM*, métricas de qualidade de imagens.

Abstract

Image steganography refers to the process of embedding a secret message into a image file without causing a noticeable visual change to anyone who has access to this image. Due to the continuous development of new steganography techniques, there is a need to develop new ways of detecting these techniques. This thesis aimed to investigate the efficiency of using Machine Learning to detect steganography in images using the Support Vector Machine technique and features by image quality measures

Keywords: Steganography, SVM, image quality measures.

Lista de figuras

Lista de abreviaturas e siglas

BMP	*Windows Bitmap*
SVM	*Support Vector Machine*
LSB	*Least Significant Bit*
PVD	*Pixel Value Differencing*
DCT	*Discrete Cosine Transform*
PNG	*Portable Network Graphics*

Sumário

1 Introdução

A necessidade de esconder informações está presente desde os primórdios da civilização humana. Técnicas de criptografia e esteganografia são utilizadas para transmitir informações de forma secreta entre um emissor e um receptor. Enquanto as técnicas de criptografia têm como objetivo impedir a leitura do conteúdo real da mensagem por invasores, a finalidade da esteganografia é esconder a própria existência da mensagem (KUMAR; POOJA, 2010).

Nos últimos anos, com a popularização do acesso à internet e aos computadores, houve também um aumento na utilização de técnicas de esteganografia digital. Tais técnicas consistem em esconder mensagens em arquivos de mídia digital, como imagens e vídeos, de forma a não causar mudança perceptível a quem tenha acesso a esse arquivo. O receptor, possuindo conhecimento prévio da existência da mensagem, é capaz de extrai-la (SHIH, 2017).

Figura 1 – Fluxograma de esteganografia

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

A Figura 1 exibe o funcionamento de um processo de esteganografia. Uma mensagem M é incorporada a um arquivo de cobertura C, resultando em um arquivo que poderá ser submetido ao processo de extração para que a mensagem seja recuperada.

Qualquer arquivo digital pode ser utilizado para esteganografia, porém arquivos que possuem um grande número de bits redundantes são mais adequados. Bits redundantes são bits que podem ser alterados sem causar mudança perceptível a humanos no conteúdo do arquivo. Isso permite a modificação de tais bits para conter a mensagem que está sendo incorporada (SHIH, 2017).

Como arquivos de imagem possuem um grande número de bits redundantes, esse é o meio mais utilizado para esteganografia. Existem técnicas para incorporar texto, áudio e até arquivo executáveis dentro de uma imagem sem que o seu conteúdo seja visualmente alterado (SILVA; CARVALHO; MARTINS, 2020).

Esteganálise é o nome dado ao processo de detecção de esteganografia. Existem dois métodos utilizados para detecção de arquivos modificados: análise visual e análise estatística. A detecção por análise visual consiste em comparar o arquivo com a sua cópia original, normalmente com auxílio de um computador para que a comparação seja feita bit a bit. Apesar

de simples, esse método não é efetivo, pois na maioria dos casos a cópia original do arquivo não está disponível para análise (SHIH, 2017).

O método de análise estatística é o mais utilizado na prática e consiste em analisar as propriedades do arquivo de forma a verificar se estão ou não dentro de um padrão. Dessa forma, a esteganografia é detectada através das alterações estatísticas causadas ao incorporar a mensagem no arquivo (SHIH, 2017).

Neste trabalho foram utilizadas técnicas de aprendizado de máquina para analisar as propriedades estatísticas de arquivos de imagem de forma a detectar a presença de uma mensagem escondida por técnicas de esteganografia.

1.1 Problema

Com o aumento na utilização de métodos de esteganografia e com o contínuo desenvolvimento de novas técnicas para isso, há a necessidade de melhorar as formas de detecção de esteganografia existentes.

Como a detecção de esteganografia utilizando análise estatística é baseada em propriedades do arquivo que fogem de um padrão, é possível utilizar técnicas de aprendizado de máquina para criar um classificador capaz de aprender esse padrão e informar se determinada imagem possui ou não uma mensagem escondida através de esteganografia.

1.2 Justificativa

Existem diversas ferramentas capazes de esconder mensagens em uma imagem utilizando esteganografia. Apesar dessa facilidade ser benéfica para usos legais, organizações criminosas e terroristas já utilizaram esteganografia para se comunicar através da internet pública sem levantar suspeiras (SHIH, 2017).

Dessa forma, o estudo de métodos de detecção automática de esteganografia é justificado através da sua importância no monitoramento de comunicações ilegais através de uma rede.

1.3 Objetivos

Objetivo Geral: Elaborar um classificador capaz de receber uma imagem como entrada e detectar se a imagem possui ou não uma mensagem incorporada através de uma técnica de esteganografia.

Objetivos Específicos:

a) Estudar conceitos e técnicas de aprendizado de máquina e de esteganografia;

b) Reunir um conjunto de dados adequado para o treinamento do modelo de classificação;

c) Definir a arquitetura e a técnica utilizada no classificador;

d) Implementar o classificador e treina-lo utilizando os dados coletados;

e) Avaliar o desempenho e a taxa de acerto do classificador;

2 Fundamentação Teórica

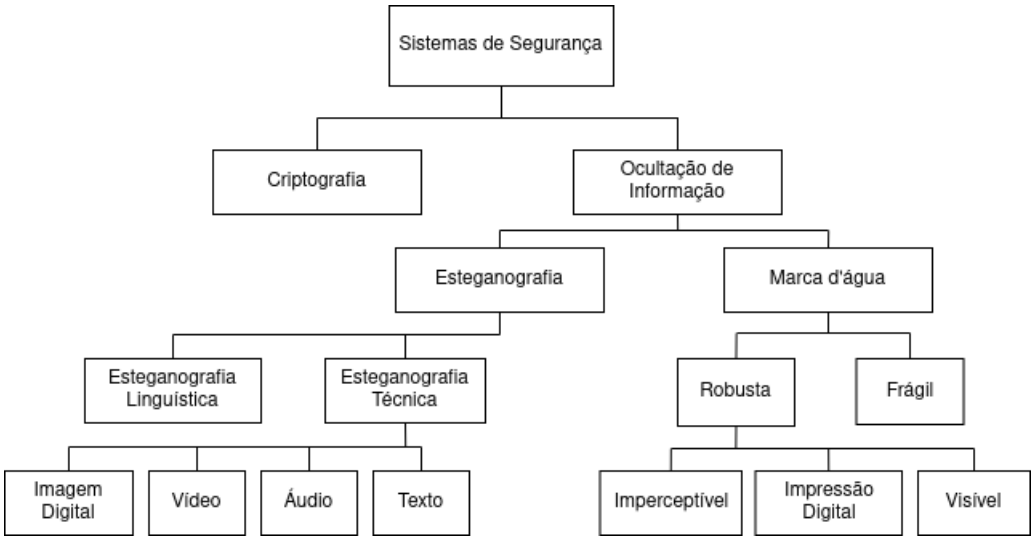
Esta seção introduz os conceitos teóricos necessários para a compreensão do trabalho.

2.1 Técnicas de Segurança da Informação

Técnicas desenvolvidas com o objetivo de estabelecer uma comunicação secreta entre duas partes existem desde a época da Grécia Antiga. Mensagens eram tatuadas na cabeça de mensageiros carecas, que esperavam até seu cabelo crescer para entregar a mensagem pessoalmente ao receptor. Ao chegar no destino, seu cabelo era cortado, revelando novamente a mensagem (SHIH, 2017). Tintas invisíveis também foram utilizadas por outras civilizações com esse objetivo. Quando aquecida, a tinta se tornava mais escura, revelando a mensagem (THAMPI, 2008).

Atualmente, com a predominância da comunicação por meios digitais, novas técnicas de segurança da informação foram desenvolvidas para atingir variados objetivos. Essas técnicas podem ser classificadas em criptografia, esteganografia e marca d'água. Uma visão geral da classificação é dada pela figura 2.

Figura 2 – Classificação dos Sistemas de Segurança da Informação



Fonte: Adaptado de Kadhim et al. (2019)

Técnicas de criptografia têm como objetivo transformar a informação em uma forma criptografada que é compreensível apenas para o emissor e para o receptor da mensagem. Essa mensagem é adequada para transmissão por meios públicos, visto que terceiros não conseguem

ter acesso à mensagem original. O processo de criptografia e decriptografia é realizado utilizando uma ou mais chaves, dependendo da técnica utilizada (KADHIM et al., 2019).

Técnicas de esteganografia são utilizadas para esconder a existência de uma mensagem durante o processo da comunicação. O objetivo é esconder a mensagem secreta dentro de uma outra mensagem pública de forma que terceiros não consigam detectar a presença de uma segunda mensagem (THAMPI, 2008).

Embora a criptografia e a esteganografia tenham o objetivo de estabelecer uma comunicação secreta entre duas partes, a definição de robustez a ataques é diferente entre as duas técnicas. Um sistema de criptografia é considerado ineficaz quando um terceiro tem acesso à mensagem original, enquanto um sistema esteganográfico é considerado ineficaz quando um terceiro consegue detectar a presença da mensagem secreta (KADHIM et al., 2019).

Técnicas de marca d'água são usadas para identificar o criador, dono, distribuidor ou o consumidor autorizado de um documento. Historicamente, marcas d'água foram utilizadas em papel para identificar uma editora e desencorajar falsificação por concorrentes. Atualmente, técnicas de marca d'água digital ganharam grande destaque como forma de inserir um padrão de bits em um meio digital para identificar o criador ou os usuários autorizados. Diferente das marcas d'água visíveis e impressas, marcas d'água digitais são projetadas para serem invisíveis aos usuários. Marcas d'água digitais também precisam ser robustas o suficiente para sobreviver a detecção, compressão e outras operações que possam ser aplicadas a um documento (SHIH, 2017).

Tanto a esteganografia digital quanto a marca d'água digital possuem o objetivo de esconder dados em um arquivo de mídia digital. No entanto, o principal objetivo da esteganografia é a imperceptibilidade, enquanto a marca d'água tem como maior objetivo a robustez. Outra diferença entre as duas técnicas está na escolha do arquivo de mídia a ser usado. Técnicas de esteganografia podem ser utilizadas com o arquivo que for mais conveniente, enquanto, no caso da marca d'água, há a necessidade de que um arquivo particular seja utilizado (KADHIM et al., 2019).

Uma comparação entre esteganografia, marca d'água e criptografia é dada pelo quadro 1.

Quadro 1 – Comparação entre esteganografia, marca d'água e criptografia.

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)

2.2 Esteganografia Digital

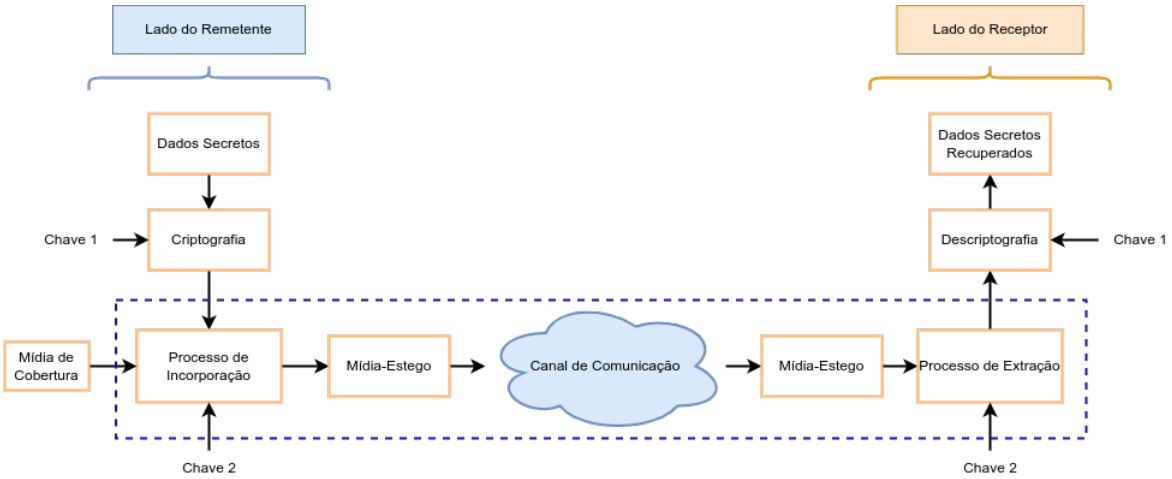
Esteganografia digital é o nome dado ao processo de esconder dados dentro de um arquivo de mídia digital, como imagens, áudios ou qualquer outro tipo de arquivo.

Na maioria dos casos o processo de esteganografia digital é realizado em imagens por terem uma grande quantidade de bits redundantes, que são bits que podem ser alterados sem causar mudança visual perceptível na imagem. Arquivos de imagem também têm a vantagem de serem facilmente compartilháveis em diversos serviços através da internet.

O processo geral de esteganografia em imagens é demonstrado pela figura 3. Dados secretos, opcionalmente criptografados, são incorporados em uma imagem de cobertura utilizando uma chave, gerando uma imagem-estego que pode ser transmitida até o receptor através de um canal público. O receptor, ao receber a imagem e conhecendo a chave que foi utilizada, consegue extrair a mensagem incorporada, realizando a decriptografia se necessário.

Existem três propriedades essenciais de qualquer sistema de esteganografia: segurança, capacidade e robustez. Segurança se refere à garantia de comunicação secreta sem levantar suspeitas de terceiros. Capacidade é a medida do número de dados que podem ser incorporados na imagem. Robustez se refere à resistência da mensagem à operações de compressão e processamento realizadas na imagem na qual ela está incorporada (WANG; WANG, 2004). Os sistemas de esteganografia buscam um balanço adequado entre essas propriedades, visto que, ao tentar melhorar uma delas, outra propriedade é prejudicada. Por exemplo, aumentar a capacidade pode reduzir a segurança ou a robustez (KADHIM et al., 2019).

Figura 3 – Diagrama do Processo de Esteganografia em Imagens



Fonte: Adaptado de Kadhim et al. (2019)

2.3 Técnicas de Esteganografia Digital em Imagens

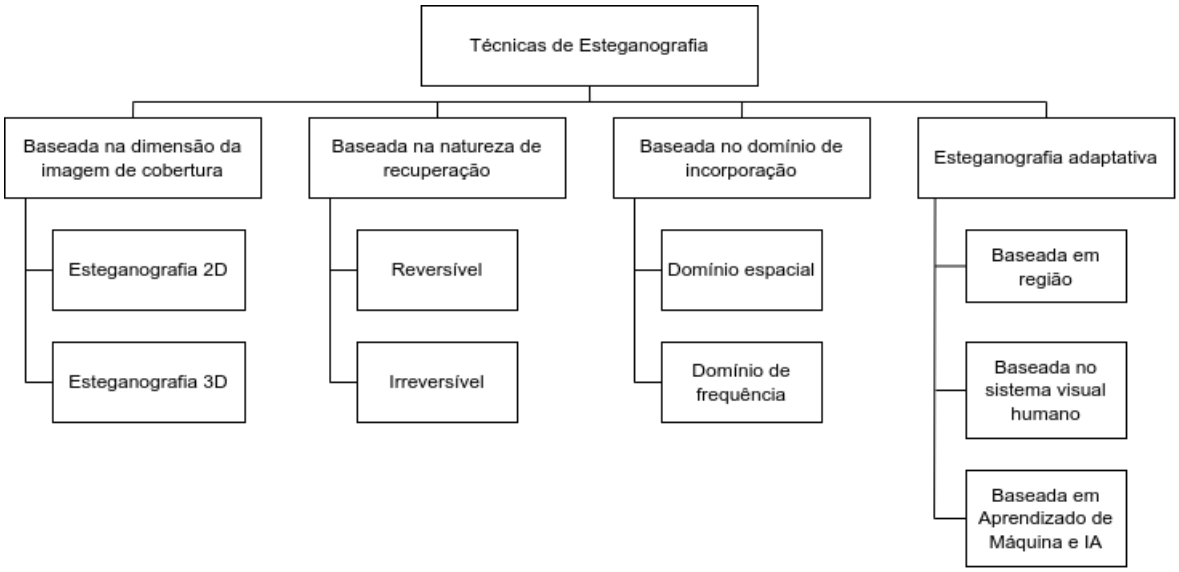
Técnicas de esteganografia podem ser classificadas quanto à dimensão da imagem de cobertura, quanto à natureza de recuperação, quanto ao domínio de incorporação, ou em esteganografia adaptativa, como demonstrado pela figura 4. A classificação mais frequente na literatura é baseada no domínio de incorporação, subdividindo-se em técnicas que utilizam o domínio espacial e técnicas que utilizam o domínio de frequência.

Técnicas classificadas como esteganografia no domínio espacial modificam os valores de intensidade dos pixels da imagem de cobertura diretamente ou indiretamente para incorporar a mensagem secreta. Essas técnicas são as mais simples e com menor complexidade de incorporação e decodificação (KADHIM et al., 2019).

As técnicas classificadas como esteganografia no domínio de frequência utilizam métodos como a transformada de Fourier e a transformada discreta de cosseno para incorporar a mensagem secreta dentro dos coficientes de frequência[1]. Esse processo, embora mais complexo, possui a vantagem de ser mais resistente a compressão, corte, redimensionamento e rotação da imagem, ou seja, são técnicas que proporcionam maior robustez (KADHIM et al., 2019).

[1] Informações sobre a transformada de Fourier e a transformada discreta de cosseno estão disponíveis em: <https://news.mit.edu/2009/explained-fourier> e <https://epubs.siam.org/doi/pdf/10.1137/S0036144598336745>

Figura 4 – Classificação das Técnicas de Esteganografia



Fonte: Adaptado de Kadhim et al. (2019)

As subseções a seguir descrevem as técnicas de esteganografia LSB, PVD e JSteg, utilizadas neste trabalho.

2.3.1 Substituição LSB

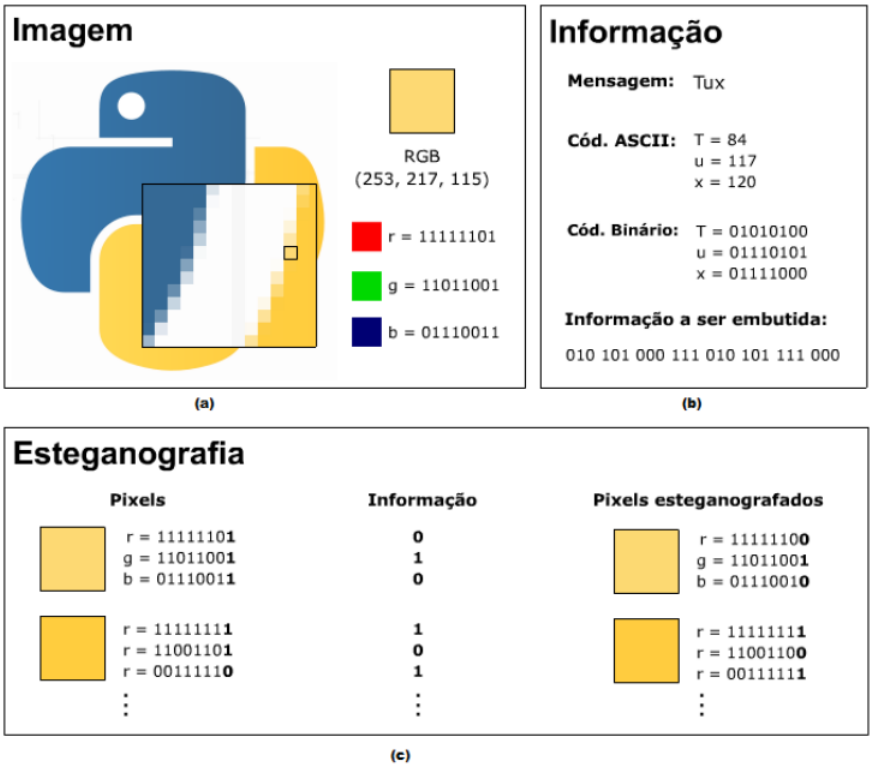
A técnica de substituição LSB (do inglês *Least Significant Bit*) é uma técnica de esteganografia baseada no domínio espacial. Ela consiste em utilizar os bits menos significativos dos pixels de uma imagem para armazenar uma mensagem secreta. No caso de uma imagem RGB de profundidade de 24 bits, cada pixel é representado por três bytes que definem a intensidade das cores vermelho, verde e azul na tonalidade do pixel. O bit menos significativo de cada um desses bytes pode ser alterado sem causar mudança visual na imagem, visto que a alteração no valor do pixel será mínima. Dessa forma, cada bit da mensagem secreta é armazenado em um desses bits menos significativos dos pixels da imagem, como ilustrado pela figura 5.

A substituição LSB é uma das técnicas de esteganografia espacial mais simples de serem entendidas e populares. Ela é eficaz quando o canal de comunicação está suscetível a ataques visuais humanos, visto que o olho humano não consegue visualizar a diferença entre a imagem original e a imagem que possui uma mensagem secreta incorporada por LSB.

Existem variações da ideia da técnica apresentada. É possível, por exemplo, definir a posição em que a incorporação da mensagem terá início na imagem. Também é possível que os bits da mensagem secreta sejam incorporados de maneira não sequencial.

Um estudo realizado em 2011 demonstrou que cerca de 70% dos softwares de este-

Figura 5 – Esteganografia LSB



Fonte: Sgursky (2015)

ganografia utilizam o algoritmo de substituição LSB ou uma de suas variações (FRIDRICH; KODOVSKỲ, 2012). Devido à sua popularidade, essa técnica ainda continua sendo ativamente estudada no campo da esteganálise.

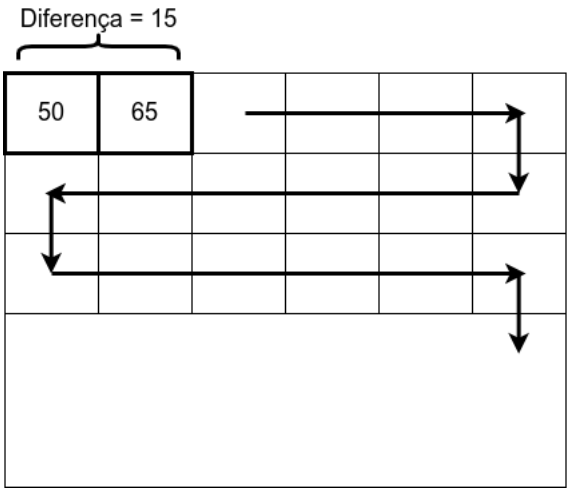
2.3.2 PVD

A técnica PVD (do inglês *Pixel Value Differencing*) é uma técnica de esteganografia baseada no domínio espacial. Ela é utilizada para esconder uma mensagem através da comparação das diferenças dos valores de dois pixels sucessivos. Esse processo é realizado dividindo a imagem 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 caso a diferença de pixels esteja dentro da faixa desejada (SHIH, 2017).

Uma das grandes vantagens dessa técnica é a possibilidade de incorporar mais bits em regiões nas quais as alterações são menos perceptíveis pelo sistema visual humano. Por exemplo, nos blocos em que o valor da diferença de valor de pixel é alto, indicando uma região de borda, é possível armazenar uma quantidade maior de bits da mensagem secreta, visto que alterações nesse tipo de região são menos perceptíveis pela visão humana (SAHU; PADHY; GANTAYAT, 2021).

O processo de esteganografia PVD é demonstrado pela figura 6, que exibe uma imagem com um grupo de pixels com valores 50 e 65, resultando em uma diferença de valor 15, que pode ser usada para definir o número de bits que será incorporado nesse grupo. Esse processo é repetido para toda a imagem, realizando a divisão em grupos de acordo com as direções demonstradas pelas flechas.

Figura 6 – Processo de esteganografia PVD



Fonte: Elaborado pelo autor

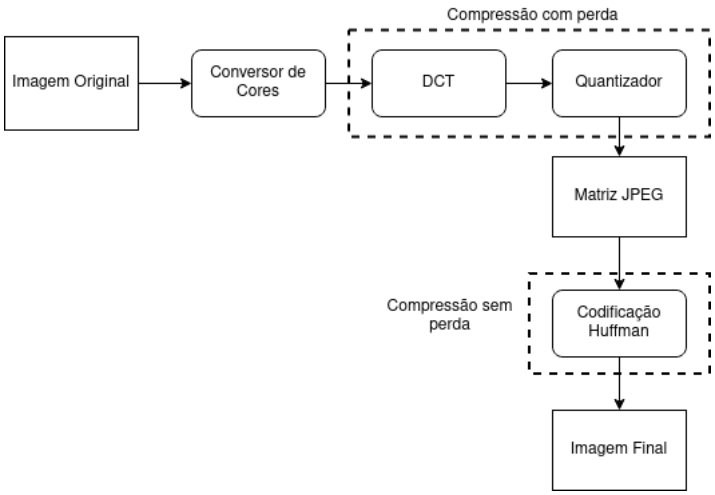
2.3.3 JSteg

JSteg é um algoritmo de esteganografia utilizado em imagens JPEG baseado no domínio de frequência. O JPEG é um formato de imagem com perda que descarta dados não essenciais ao sistema visual humano com o objetivo de reduzir o tamanho do arquivo da imagem. Devido a essa característica, técnicas de esteganografia no domínio espacial não podem ser realizadas diretamente nos arquivos desse formato (SCHAATHUN, 2012). Uma visão geral do processo de compressão JPEG e dado pela figura 7.

Durante o processo de compressão com perda, cada componente de cor é dividido em blocos de tamanho 8x8. Cada bloco de 8x8 pixels é mapeado através de uma Transformada de Cosseno bidimensional, produzindo um bloco 8x8 de coeficientes DCT (SCHAATHUN, 2012).

O algoritmo de esteganografia JSteg consiste em realizar a substituição de bits menos significativos dos coeficientes DCT de forma a não causar mudança perceptível quando a imagem é convertida para o domínio espacial (SHEISI; MESGARIAN; RAHMANI, 2012).

Figura 7 – Processo de compressão JPEG



Fonte: Adaptado de Schaathun (2012)

2.4 Métricas de Qualidade da Imagem

Métricas de qualidade da imagem são um conjunto de métricas que possuem o objetivo de comparar uma imagem com outra imagem distorcida ou adulterada e avaliar a diferença entre elas (SHIH, 2017).

Foi sugerido por Sgursky (2015) a utilização de tais métricas como características para detecção de esteganografia ao realizar o cálculo utilizando a imagem a ser analisada e essa mesma imagem convertida para outro formato.

As equações a seguir descrevem o cálculo das métricas entre duas imagens de tamanho M x N pixels. *F(j,k)* se refere ao valor do pixel na linha *j* e coluna *k* da primeira imagem, enquanto *G(j,k)* se refere ao pixel de mesma posição na segunda imagem (SHIH, 2017).

- Distância Média

$$\sum_{j=1}^{M}\sum_{k=1}^{N}(F(j,k)-G(j,k))/MN \tag{2.1}$$

- Distância Euclidiana

$$\frac{1}{MN}(\sum_{j=1}^{M}\sum_{k=1}^{N}(F(j,k)-G(j,k))^2)^{1/2} \tag{2.2}$$

- Conteúdo Estrutural

$$\sum_{j=1}^{M}\sum_{k=1}^{N}F(j,k)^2/\sum_{j=1}^{M}\sum_{k=1}^{N}G(j,k)^2 \tag{2.3}$$

- Fidelidade da Imagem

$$1-(\sum_{j=1}^{M}\sum_{k=1}^{N}(F(j,k)-G(j,k))^2/\sum_{j=1}^{M}\sum_{k=1}^{N}G(j,k)^2) \tag{2.4}$$

- Correlação Cruzada Normalizada

$$\sum_{j=1}^{M}\sum_{k=1}^{N} F(j,k)G(j,k) / \sum_{j=1}^{M}\sum_{k=1}^{N} F(j,k)^2 \tag{2.5}$$

- Erro Médio Quadrático Normal

$$\sum_{j=1}^{M}\sum_{k=1}^{N} (F(j,k) - G(j,k))^2 / \sum_{j=1}^{M}\sum_{k=1}^{N} F(j,k)^2 \tag{2.6}$$

- Erro Médio Quadrático Mínimo

$$\sum_{j=1}^{M-1}\sum_{k=2}^{N-1} (F(j,k) - G(j,k))^2 / \sum_{j=1}^{M-1}\sum_{k=2}^{N-1} O(F(j,k))^2 \tag{2.7}$$

$$O(F(j,k)) = F(j+1,k) + G(j-1,k) + F(j,k+1) + F(j,k-1) - 4F(j,k) \tag{2.8}$$

- Pico do Erro Médio Quadrático

$$\frac{1}{MN}\sum_{j=1}^{M}\sum_{k=1}^{N} [F(j,k) - G(j,k)]^2 / \{max_{j,k}[F(j,k)]\}^2 \tag{2.9}$$

- Sinal de Pico-Ruído

$$20 \times log_{10}\{255 / \{\sum_{j=1}^{M}\sum_{k=1}^{N} [F(j,k) - G(j,k)]^2\}^{1/2}\} \tag{2.10}$$

2.5 Aprendizado de Máquina

Aprendizado de máquina é a área da Inteligência Artificial que estuda algoritmos e técnicas que tornam um sistema capaz de aprender com dados, possibilitando, por exemplo, reconhecer o dígito numérico que está presente em uma imagem ou classificar e-mails como sendo spam ou não.

O processo de resolução de um problema usando aprendizado de máquina envolve a formação de um conjunto de dados e a construção de um modelo estatístico baseado nesse conjunto de dados (BURKOV, 2019). Nesse conjunto, cada elemento é considerado uma amostra e é formado por um vetor de valores numéricos, denominados características, que é o que diferencia uma amostra de outra. Cada amostra também pode ter um rótulo associado a ela indicando a classe que ela representa.

Os algoritmos de aprendizado de máquina dividem-se de acordo com o problema que buscam resolver, podendo ser dos seguintes tipos:

- **Classificação**: nesse caso, o objetivo é identificar a qual classe cada amostra pertence a partir de uma lista de classes predefinida.

- **Regressão**: na regressão, o objetivo é predizer um valor contínuo ou uma variável númerica associada à amostra.
- **Clusterização**: nesse caso, busca-se agrupar amostras de forma que as amostras em um mesmo grupo tenham um nível de semelhança maior entre si em comparação com o resto das amostras.
- **Otimização**: na otimização, é realizada uma série de comparações de soluções possíveis até que uma solução ótima ou satisfatória seja encontrada.

Dentro do processo de aprendizado de máquina, o aprendizado pode ser de diversos tipos, entre eles supervisionado, não supervisionado, semi supervisionado e por reforço, descritos a seguir:

- **Aprendizado supervisionado**: nesse caso, o conjunto de dados é rotulado. O objetivo é produzir um modelo que recebe como entrada um vetor de características de um elemento e gera como saída um dado que permite saber o rótulo do elemento.
- **Aprendizado não supervisionado**: o conjunto de dados não é rotulado. O objetivo é transformar um vetor de características em outro vetor ou em um valor que pode ser usado para resolver um problema prático.
- **Aprendizado semi supervisionado**: o conjunto de dados contém elementos rotulados e elementos não rotulados. O objetivo é o mesmo do aprendizado supervisionado. A ideia é possibiliar a criação de um modelo melhor através do uso de elementos não rotulados.
- **Aprendizado por reforço**: nesse caso, o algoritmo está constantemente rodando em um ambiente e é capaz de percebê-lo através de um vetor de características. Várias ações podem ser executadas em cada estado, que trazem diferentes recompensas e podem alterar o estado para outro. O objetivo é encontrar a melhor ação para cada estado, ou seja, a ação que maximiza a recompensa.

2.5.1 Técnicas de Aprendizado Supervisionado

Técnicas de aprendizado supervisionado normalmente são utilizadas em problemas de classificação e regressão. Exemplos de algoritmos dessa categoria são descritos a seguir:

- **KNN**: a técnica K Vizinhos Mais Próximos (do inglês *K-Nearest Neighbors*) realiza a classificação com base nos K vizinhos mais próximos de uma amostra, definindo como classe da amostra a classe predominante em seus vizinhos.
- **SVM**: o SVM (*Support Vector Machine*) é um algoritmo capaz de encontrar um hiperplano que separa as amostras de uma classe das amostras de outra classe, podendo utilizar esse hiperplano para classificar novas amostras.

- **Árvore de Decisão**: essa técnica utiliza a estrutura de dados denominada árvore para classificar amostras, partindo do nó raiz até o nó folha, sendo que em cada ramificação a decisão do caminho a ser seguido é tomada analisando uma característica da amostra.
- **Redes Neurais**: os sistemas de redes neurais artificiais simulam a forma como as redes neurais biológicas funcionam, utilizando uma série de camadas compostas por nós ou neurônios, sendo uma camada de entrada, uma ou mais camadas intermediárias e uma camada de saída. Cada neurônio artificial pode se conectar a outro utilizando um valor numérico de peso.

A subseção a seguir descreve com mais detalhes a técnica SVM, utilizada neste trabalho.

2.5.1.1 *Support Vector Machine* (SVM)

Support Vector Machine (SVM) é um algoritmo de aprendizado supervisionado. Esse algoritmo utiliza como entrada um conjunto de elementos rotulados nos quais cada rótulo representa a classe do elemento, podendo ser 1 para a classe positiva e -1 para a classe negativa.

Cada elemento é representado por um ponto em um espaço multi-dimensional. O algoritmo, então, encontra um hiperplano capaz de separar elementos que possuem classe positiva dos elementos que possuem classe negativa. Essa fronteira que separa os elementos de cada classe é denominada fronteira de decisão.

Em conjuntos linearmente separáveis, apesar de existirem infinitos hiperplanos que separam os elementos de cada classe, o SVM é capaz de encontrar o hiperplano que maximiza a margem entre os elementos mais próximos de cada classe, o que contribui para uma melhor classificação de dados.

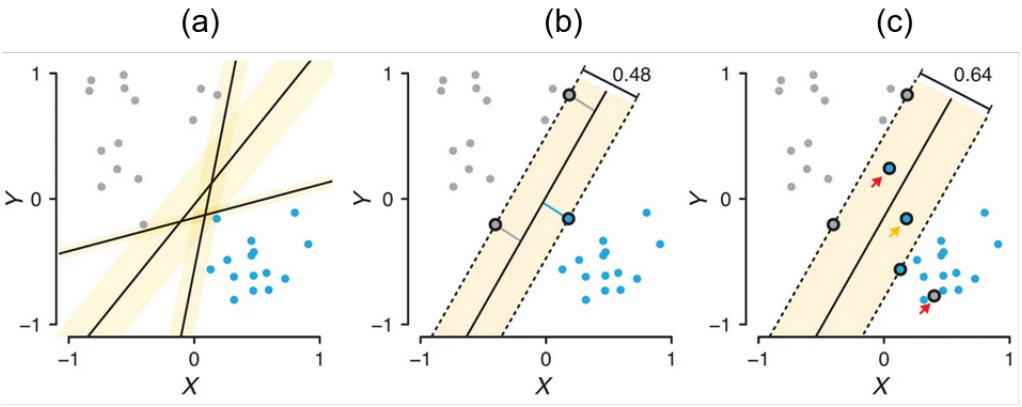
Esse processo é ilustrado pela figura 8, demostrando a classificação de amostras em duas classes possíveis, indicadas pelas cores cinza e azul. Em *a*, é possível ver alguns dos infinitos hiperplanos que separam as amostras de cada classe. Em *b*, após a execução do SVM, é exibido o hiperplano que maximiza a margem entre as classes. Em *c*, é demonstrado qual seria o resultado caso algumas amostras fossem adicionadas nos locais indicados, tornando o conjunto não linearmente separável, o que geraria erros de classificação.

Nos casos em que o conjunto de dados não é linearmente separável, é possível utilizar funções Kernel de modo a mapear os dados para um novo espaço de dimensão maior, no qual há a possibilidade dos dados serem separáveis por um hiperplano (CHAUHAN; DAHIYA; SHARMA, 2019). Esse processo é ilustrado pela figura 9.

2.5.2 Técnicas de Aprendizado Não Supervisionado

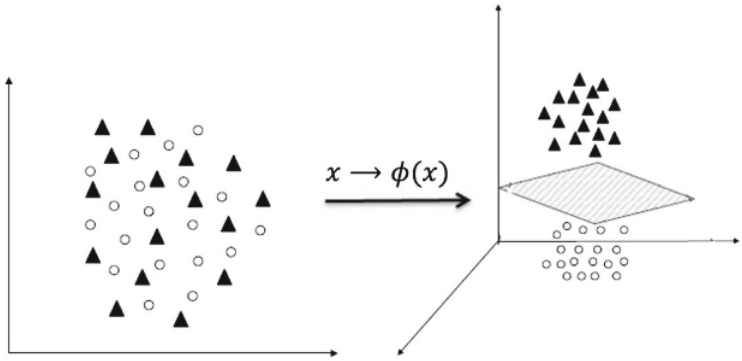
Em conjuntos de dados não rotulados, normalmente são utilizadas técnicas de aprendizado não supervisionado baseadas em clusterização. Exemplos de algoritmos dessa categoria

Figura 8 – Exemplo de hiperplano encontrado pelo SVM



Fonte: Bzdok, Krzywinski e Altman (2018)

Figura 9 – Demonstração do uso de uma função Kernel



Fonte: Chauhan, Dahiya e Sharma (2019)

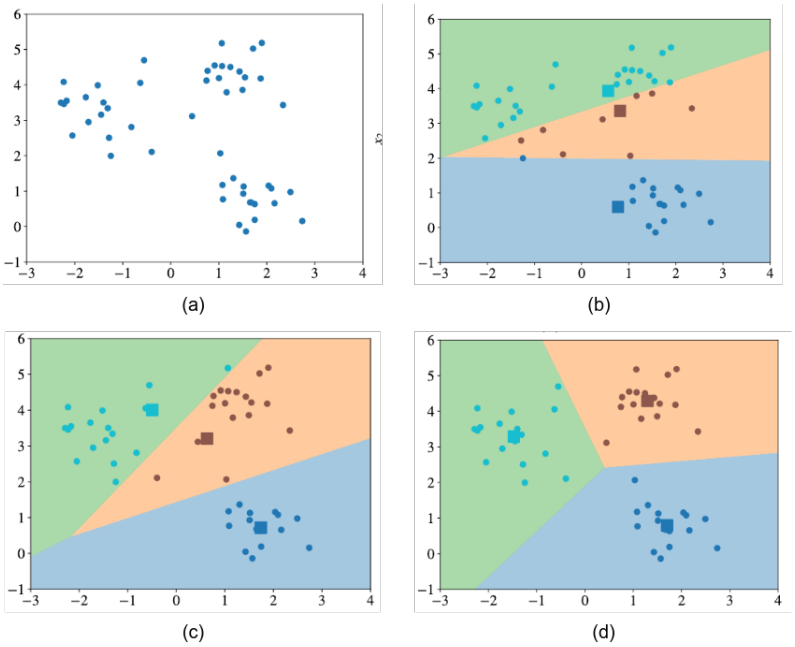
são:

- **K-means**: essa técnica particiona as amostras em k grupos, utilizando para isso k centroides inicialmente em posições aleatórias, sendo cada elemento associado à centroide mais próxima. Em cada iteração do algoritmo, cada centroide é reposicionada de acordo com a média das amostras associada a ela.
- **DBSCAN**: o DBSCAN (*Density-based spatial clustering of applications with noise*) realiza o agrupamento sem a necessidade de uma predefinição do número de clusters. Isso é feito detectando regiões nas quais existe um grande número de amostras próximas, ou seja, regiões densas, separadas por regiões relativamente vazias.

A figura 10 mostra um exemplo de clusterização por K-means utilizando três grupos. O conjunto de amostras em seu estado original, demonstrado em *a*, é submetido a diversas

iterações do algoritmo de modo a reposicionar os centroides de cada grupo para melhor clusterização. O resultado após 1 iteração, 3 iterações e 5 iterações é mostrado em *b*, *c* e *d*, respectivamente.

Figura 10 – Exemplo de clusterização utilizando K-means



Fonte: Adaptado de Burkov (2019)

2.5.3 Métricas de avaliação

Uma das formas de avaliar o desempenho de um classificador é utilizando uma matriz de confusão. Essa é uma tabela que resume o sucesso do modelo ao prever a classe de cada elemento do conjunto de testes. Um dos eixos da matriz de confusão é a classe prevista pelo modelo, enquanto o outro eixo é a classe real do elemento (BURKOV, 2019).

A matriz de confusão fornece quatro valores que são úteis para o cálculo de outras métricas:

- **Verdadeiros positivos (VP)**: número de casos em que o modelo preveu o rótulo como positivo e ele era realmente positivo
- **Falsos negativos (FN)**: número de casos em que o modelo preveu o rótulo como negativo, porém ele era positivo
- **Falsos positivos (FP)**: número de casos em que o modelo preveu o rótulo como positivo, porém ele era negativo

- **Verdadeiros negativos (VN)**: número de casos em que o modelo preveu o rótulo como negativo e ele era realmente negativo

A partir desses valores é possível calcular as métricas de precisão, recall e acurácia do modelo a partir das equações 2.11, 2.12 e 2.13, respectivamente.

$$Precisão = \frac{VP}{VP + FP} \tag{2.11}$$

$$Recall = \frac{VP}{VP + FN} \tag{2.12}$$

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.13}$$

A precisão é a taxa de predições positivas corretas em relação ao número total de predições positivas. Recall é a taxa de predições positivas corretas em comparação com o total de elementos positivos. Acurácia é a taxa de elementos classificados corretamente em relação ao número total de elementos (BURKOV, 2019).

3 Desenvolvimento

3.1 Estrutura Geral

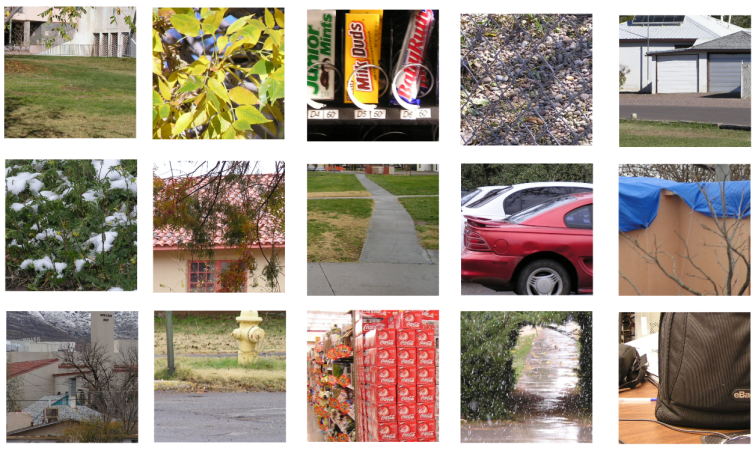
As técnicas de esteganografia selecionadas para a realização do trabalho foram Substituição LSB, PVD e JSteg, escolhidas de acordo com os critérios de popularidade, simplicidade, disponibilidade de implementações de referência e relevância dentro da área de esteganálise.

Foi utilizada a abordagem de construção de uma classificador binário específico para cada técnica. A técnica de aprendizado de máquina escolhida foi o SVM devido a ser a técnica mais utilizada em trabalhos relacionados.

3.2 Base de dados

Para esse trabalho foi utilizada uma base de imagens fornecida por Liu, Cooper e Zhou (2013) que contém 5150 imagens coloridas não compactadas de tamanho 256 x 256 pixels no formato BMP[1]. Alguns exemplos de imagens contidas nessa base de dados podem ser vistos na figura 11.

Figura 11 – Exemplo de imagens contidas no banco de imagens utilizado no trabalho



Fonte: Elaborado pelo autor

[1] Disponível em: <https://www.shsu.edu/qxl005/New/Downloads/index.html>

3.3 Ferramentas Utilizadas

Todo o código de construçao do classificador e análise dos resultados foi desenvolvido utilizando a linguagem de programação Python, escolhida por ser a linguagem mais utilizada em aplicações de aprendizado de máquina. As técnicas de esteganografia foram implementadas nas linguagens Python e Go. O desenvolvimento foi feito através do editor de código *Visual Studio Code*[2] e executado dentro da plataforma *Google Colaboratory*[3].

As bibliotecas *NumPy*[4] e *Pandas*[5] foram utilizadas para manipulação numérica dos dados ao longo do trabalho. A implementação de técnicas de esteganografia e das métricas usadas para treinamento foi feita utilizando as bibliotecas *Pillow*[6] e *OpenCV*[7].

Para a construção dos classificadores, foi utilizada a biblioteca *scikit-learn*[8], que possui a implementação de algoritmos de aprendizado de máquina e outras funções de suporte ao processo de treinamento e teste do modelo. A biblioteca *Matplotlib*[9] foi usada para gerar gráficos e tabelas para visualização dos resultados.

3.4 Execução do Projeto

O primeiro passo realizado foi a conversão da base de imagens, originalmente no formato BMP, para os formatos PNG e JPEG. As imagens em formato PNG foram usadas para construir os conjuntos de treinamento e teste relativos às técnicas LSB e PVD, enquanto a base de imagens em JPEG foi utilizada para a técnica JSteg.

Após isso, foi realizada a implementação das técnicas e a construção dos conjuntos de treinamento e teste de cada técnica. Para isso, foi realizado o processo de esteganografia em 50% das 5150 imagens, resultando em um conjunto de 2575 imagens originais e 2575 imagens contendo uma mensagem escondida através da técnica de esteganografia em questão.

O tamanho da mensagem escondida foi definido de forma diferente de acordo com cada técnica. Para a técnica LSB, foi decidido construir um conjunto de imagens para diversos níveis de capacidade de forma a poder avaliar o desempenho do classificador e das métricas de acordo com o aumento do tamanho da mensagem. Dessa forma, foram construídos conjuntos de imagens aplicando a técnica LSB com 10%, 25%, 50%, 75% e 100% de capacidade. O cálculo de capacidade foi feito considerando o valor de 3 bits por pixel da imagem.

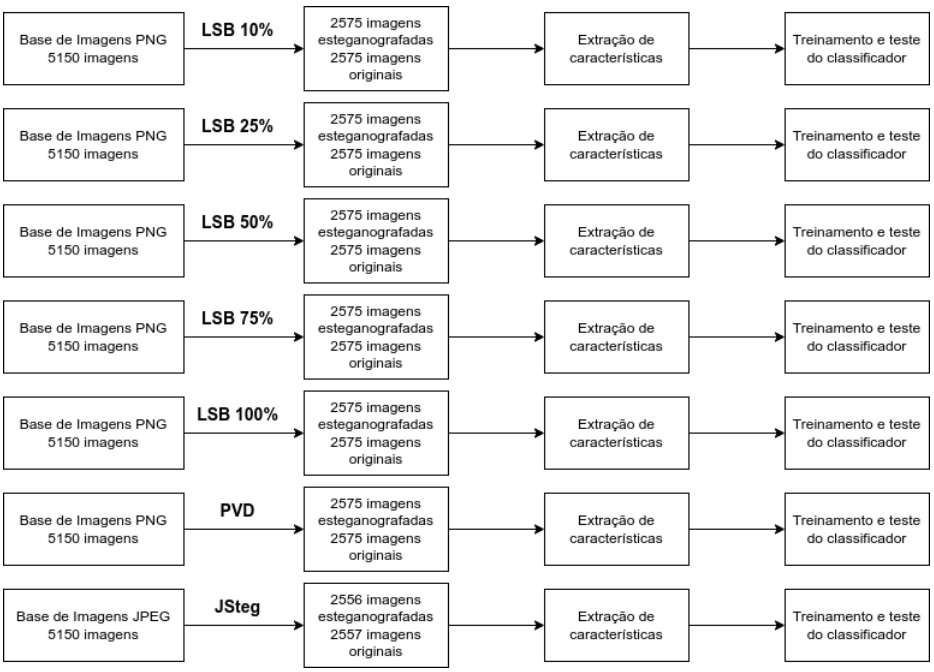
O conjunto de imagens da técnica PVD foi feito utilizando o valor fixo de 1220 caracteres

[2] Disponível em: <https://code.visualstudio.com>
[3] Disponível em: <https://colab.research.google.com>
[4] Disponível em: <https://numpy.org/>
[5] Disponível em: <https://pandas.pydata.org/>
[6] Disponível em: <https://python-pillow.org/>
[7] Disponível em: <https://opencv.org/>
[8] Disponível em: <https://scikit-learn.org>
[9] Disponível em: <https://matplotlib.org/>

por imagem. Para o JSteg, foram utilizadas mensagens de 500 caracteres para arquivos maiores que 10 KB, 100 caracteres para arquivos de tamanho entre 4 KB e 10 KB, 50 caracteres para arquivos de tamanho entre 3 KB e 4 KB e 10 caracteres para os arquivos restantes. Esses valores foram definidos empiricamente como capacidades adequadas através de testes na base de dados. 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.

A figura 12 demonstra o processo de formação dos conjuntos de imagens mencionado.

Figura 12 – Metodologia da formação dos conjuntos de imagem de cada técnica



Fonte: Elaborado pelo autor

Após formados os conjuntos de imagem de cada técnica, foi realizado o processo de extração de características. Para isso foram utilizadas as métricas de qualidade da imagem conforme descrito na seção 2.4. No caso das técnicas LSB e PVD, o cálculo foi realizado convertendo cada imagem para o formato JPEG e realizando a comparação entre a imagem PNG original e a imagem convertida. No caso do JSteg, foi realizada a comparação entre a imagem original e a mesma imagem com um filtro gaussiano aplicado, método sugerido por Schaathun (2012).

Com as caraterísticas disponíveis, a etapa final de cada técnica consistiu no treinamento e teste do classificador. O conjunto de imagens de cada técnica foi dividido em 70% para treinamento e 30% para teste, valores escolhidos por serem muito utilizados em aplicações de aprendizado de máquina.

4 Resultados

Os valores de acurácia, precisão e recall obtidos nos classificadores, assim como o número de elementos do conjunto de teste em cada caso, são exibidos na tabela 1.

Tabela 1 – Acurácia, precisão e recall dos classificadores

técnica	elementos	acurácia	precisão	recall
LSB 10%	1545	53.33%	52.56%	55.55%
LSB 25%	1545	57.22%	55.23%	64.54%
LSB 50%	1545	67.77%	65.54%	73.76%
LSB 75%	1545	72.56%	70.96%	76.29%
LSB 100%	1545	79.16%	76.98%	82.53%
DVP	1545	59.42%	60.50%	56.00%
JSteg	1534	52.02%	54.66%	32.35%

Fonte: Elaborado pelo autor

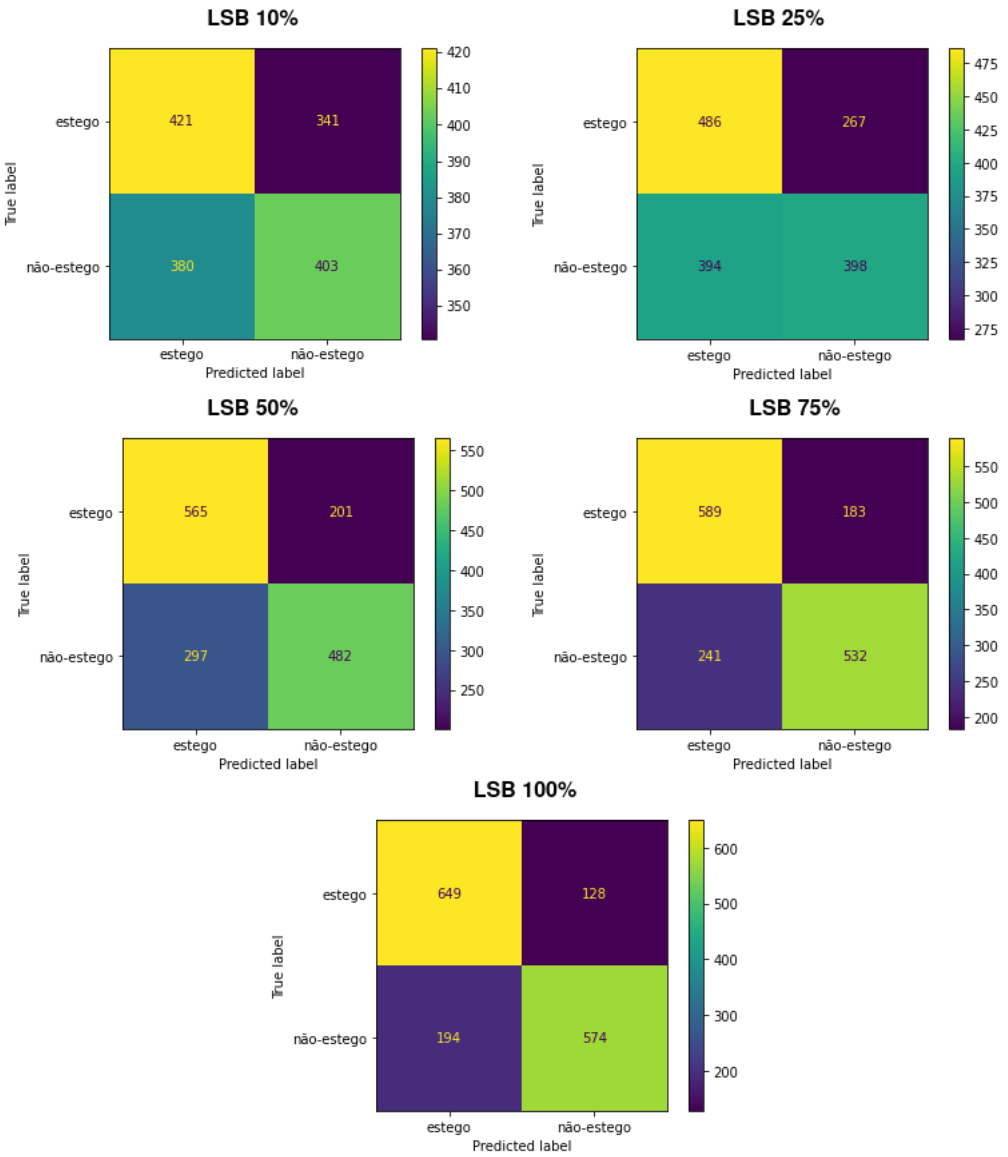
Analisando os resultados do classificador LSB, é possível observar que todas as três métricas crescem à medida que o tamanho da mensagem aumenta, indicando que as métricas de qualidade da imagem utilizadas como características refletem a presença ou não de esteganografia.

O classificador LSB demonstrou ser eficiente nos casos em que existe uma grande quantidade de informação incorporada na imagem, porém perde sua eficiência nos casos em que a informação incorporada é mínima. A matriz de confusão dos classificadores LSB é exibida na figura 13.

Os classificadores das técnicas PVD e JSteg demonstraram uma eficácia muito menor em comparação ao melhor caso do LSB. Isso provavelmente se deve à menor capacidade dessas técnicas, em especial do JSteg, que levaram à necessidade de incorporar uma quantidade mínima de informação nas imagens da base de dados utilizada. As matrizes de confusão dos dois classificadores são exibidas nas figuras 14 e 15.

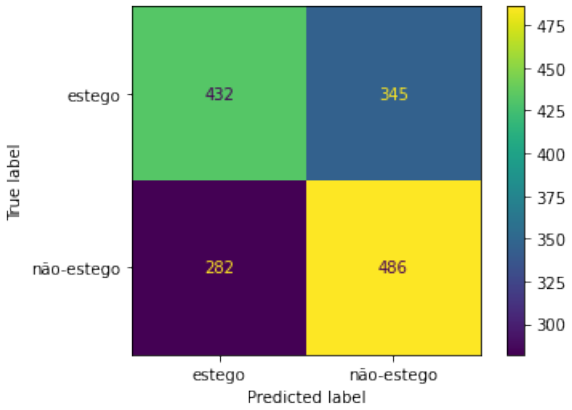
Em comparação com trabalhos similares, os resultados do LSB foram melhores que os obtidos por Schaathun (2012) ao analisar a acurácia de detecção LSB utilizando métricas de qualidade de imagem. O referido trabalho obteve acurácia de 62% para LSB utilizando 40% de capacidade e acurácia de 71.7% para LSB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

Figura 13 – Matrizes de confusão dos classificadores LSB



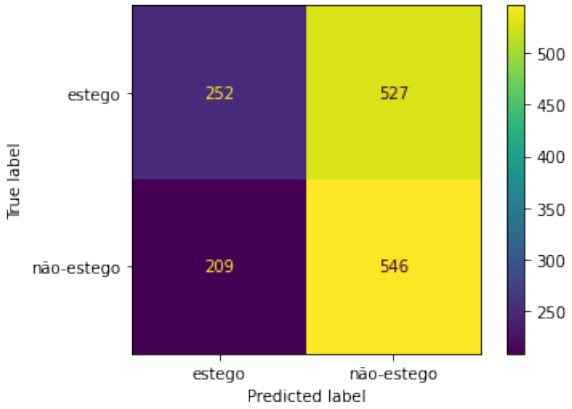
Fonte: Elaborado pelo autor

Figura 14 – Matriz de confusão do classificador PVD



Fonte: Elaborado pelo autor

Figura 15 – Matriz de confusão do classificador JSteg



Fonte: Elaborado pelo autor

5 Conclusão

As áreas de esteganografia e esteganálise continuam sendo um campo de pesquisa ativo. Enquanto novas técnicas de esteganografia são idealizadas, há a necessidade de novas formas de detectar arquivos de mídia que possuem informação secreta incorporada. Neste trabalho, foi avaliada a abordagem de uso de aprendizado de máquina para detecção de esteganografia em arquivos de imagem.

Foi realizado um estudo bibliográfico com objetivo de selecionar técnicas de esteganografia e métricas que possam ser utilizadas para treinar um classificador capaz de detectar esteganografia. 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 em uma imagem por esteganografia é grande, porém a eficiência é reduzida nos casos em que a informação é mínima.

Para trabalhos futuros utilizando essa abordagem, incentiva-se a análise da eficácia de detecção utilizando outras técnicas de esteganografia e de aprendizado de máquina. Também é interessante a investigação de métricas e características que possam detectar pequenas informações incorporadas em uma imagem.

Referências

BURKOV, A. *The hundred-page machine learning book*. [S.l.]: Andriy Burkov Quebec City, QC, Canada, 2019. v. 1.

BZDOK, D.; KRZYWINSKI, M.; ALTMAN, N. Machine learning: supervised methods. *Nature methods*, NIH Public Access, v. 15, n. 1, p. 5, 2018.

CHAUHAN, V. K.; DAHIYA, K.; SHARMA, A. Problem formulations and solvers in linear svm: a review. *Artificial Intelligence Review*, Springer, v. 52, n. 2, p. 803–855, 2019.

FRIDRICH, J.; KODOVSKỲ, J. Steganalysis of lsb replacement using parity-aware features. In: SPRINGER. *International Workshop on Information Hiding*. [S.l.], 2012. p. 31–45.

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.

KUMAR, A.; POOJA, K. Steganography-a data hiding technique. *International Journal of Computer Applications*, Citeseer, v. 9, n. 7, p. 19–23, 2010.

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.

SAHU, M.; PADHY, N.; GANTAYAT, S. S. Multi-directional pvd steganography avoiding pdh and boundary issue. *Journal of King Saud University-Computer and Information Sciences*, Elsevier, 2021.

SCHAATHUN, H. G. *Machine learning in image steganalysis*. [S.l.]: Wiley Online Library, 2012.

SGURSKY, L. F. F. Análise e implementação de técnicas de esteganografia. 2015.

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.

SHIH, F. Y. *Digital watermarking and steganography: fundamentals and techniques*. [S.l.]: CRC press, 2017.

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.

THAMPI, S. M. Information hiding techniques: a tutorial review. *arXiv preprint arXiv:0802.3746*, 2008.

WANG, H.; WANG, S. Cyber warfare: steganography vs. steganalysis. *Communications of the ACM*, ACM New York, NY, USA, v. 47, n. 10, p. 76–82, 2004.