UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
CURSO DE ENGENHARIA DE COMPUTAÇÃO

EDUARDO RODRIGO DE OLIVEIRA

ANÁLISE DE TÉCNICAS DE AGRUPAMENTOS PARA CLASSIFICAÇÃO DE SEMENTES DE SOJA

TRABALHO DE CONCLUSÃO DE CURSO

CORNÉLIO PROCÓPIO
2020

EDUARDO RODRIGO DE OLIVEIRA

ANÁLISE DE TÉCNICAS DE AGRUPAMENTOS PARA CLASSIFICAÇÃO DE SEMENTES DE SOJA

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia de Computação da Universidade Tecnológica Federal do Paraná, como requisito parcial para a obtenção do título de Bacharel.

Orientador: Priscila Tiemi Maeda Saito
Universidade Tecnológica Federal do Paraná

CORNÉLIO PROCÓPIO
2020

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Cornélio Procópio
Diretoria de Graduação e Educação Profissional
Departamento Acadêmico de Computação
Engenharia de Computação

TERMO DE APROVAÇÃO

Análise de Técnicas de Agrupamentos para Classificação de Sementes de Soja

por

Eduardo Rodrigo de Oliveira

Este Trabalho de Conclusão de Curso de graduação foi julgado adequado para obtenção do Título de Bacharel em Engenharia de Computação e aprovado em sua forma final pelo Programa de Graduação em Engenharia de Computação da Universidade Tecnológica Federal do Paraná.

Cornélio Procópio, 01/09/2020

_____

Prof.ª Dra. Priscila Tiemi Maeda Saito

_____

Prof. Dr. Cléber Gimenez Corrêa

_____

Prof. Dr. Pedro Henrique Bugatti

"A Folha de Aprovação assinada encontra-se na Coordenação do Curso"

Aos amigos e família.

AGRADECIMENTOS

A Professora Priscila T. M. Saito, pela orientação.
A Universidade Tecnológica Federal do Paraná, onde faço a graduação.
Aos amigos e familiares, que me acompanharam nessa jornada.

O homem não teria alcançado o possível se, repetidas vezes, não tivesse tentado o impossível. (WEBER, Maximillion).

RESUMO

OLIVEIRA, Eduardo. Análise de Técnicas de Agrupamentos para Classificação de Sementes de Soja. 2020. 55 f. Trabalho de Conclusão de Curso – Curso de Engenharia de Computação, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2020.

Considerado o principal produto da agricultura brasileira, a soja é o quarto grão mais cultivado no mundo e sua produção tende a aumentar. Devido a este grande mercado, a garantia de qualidade do produto torna-se um fator indispensável para empresas que querem se manter competitivas. Maneiras de validar a qualidade e adquirir informações sobre o plantio são os testes de vigor, como o teste de tetrazólio que traz informações de danos ocasionados por umidade, percevejo ou dano mecânico. Entretanto, a verificação do tipo de dano e sua gravidade são realizadas, uma a uma e, visualmente por um analista, ou seja, além de um processo demorado é susceptível ao erro, visto que é um trabalho massante e cansativo. Propostas envolvendo diferentes abordagens de aprendizado supervisionado, incluindo estratégias de aprendizado ativo, já foram utilizadas e trouxeram resultados significativos. Dessa forma, o objetivo deste trabalho é analisar o desempenho de técnicas não supervisionadas para a classificação de sementes de soja. Para tanto, foi realizada uma avaliação experimental extensiva, considerando (9) diferentes algoritmos de agrupamento (entre eles particionais, hierárquicos e baseados em densidade) aplicadas a (5) conjuntos de imagens de sementes de soja submetidas ao teste de tetrazólio, incluindo diferentes danos e/ou seus respectivos níveis. Para a descrição de tais imagens foram considerados (18) extratores de características tradicionais. Para validação foram consideradas (4) métricas (acurácia, FOWLKES, DAVIES e CALINSKI) e duas técnicas de redução de dimensionalidade (PCA e TSNE). A partir dos resultados obtidos, pode-se observar que o presente trabalho apresenta contribuições significativas, dado que possibilita identificar os descritores e algoritmos de agrupamento a serem utilizados como pré-processamento em outras abordagens de aprendizado, acelerando e melhorando o processo de classificação.
Palavras-chave: Semente de soja. Agrupamento. Aprendizado de Máquina. Teste de tetrazólio.

ABSTRACT

OLIVEIRA, Eduardo. Analysis of Clustering Techniques for Classification of Soybean Seeds. 2020. 55 f. Trabalho de Conclusão de Curso – Curso de Engenharia de Computação, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2020.

Soy is the main product of brasilian agriculture and the fourth most cultivated bean in the world. Since the cultivation of soy tends to increase and due to this large market, the garantee of the product quality is an undispensable factor for enterprises to stay competitive. To acquire information and evaluate the quality of soy planting, industries perform vigor tests. The tetrazolium test, for example, provides information about moisture damage, bedbugs or mechanical damage. However, the verification of the damage reason and it's severity are done by an analist, one by one. Since this is a massive and exausting work, it is suceptible to mistakes. Proposals involving different supervised learning approaches, including active learning strategies have already been used and brought significant results. Therefore, this paper analyzes the performance of non-supervised techniques for classifying soybeans. A extensive experimental evaluation was realized, considering (9) different clustering algorithms (partitional, hierarquichal and density based) applied to (5) image datasets of soybeans seeds submitted to the tetrazolium test, including different damages and/or their levels. To describe those images, (18) extractors of traditional features were considered. (4) metrics (accuracy, FOWLKES, DAVIES, and CALINSKI) and two dimensionality reduction techniques (PCA and TSNE) were considered for validation. Results show that this paper presents important contributions, since it makes possible to identify descriptors and clustering algorithms that shall be used as pre-processing in other learning processes, accelerating and improving the classification process.

Keywords: Soybean. Machine Learning. Clustering.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACC	*Auto Color Correlogram*
AGNES	*Agglomerative Nesting*
BIC	*Border/Interior Pixel Classification*
BIRCH	*Balanced Interactive Reducing and Clustering using Hierarchies*
CEDD	*Color and Edge Directivity Descriptor*
CLARA	*Clustering Large Applications*
CLARANS	*Clustering Large Applications based on Randomized Search*
CURE	*Clustering using Representatives*
DBSCAN	*Density Based Spatial Clustering of Applications with Noise*
DENCLUE	*Density Clustering*
DIANA	*Divisive Analysis*
FCM	*Fuzzy-c-Means*
FCTH	*Fuzzy Color and Texture Histogram*
GCH	*Global Color Histogram*
JCD	*Join Composite Descriptor*
LBP	*Local Binary Pattern*
LCH	*Local Color Histogram*
MPO	Medidas de Primeira Ordem
MPOC	Medidas de Primeira Ordem Color
OPF	*Optimum-Path Forest*
OPTICS	*Ordering Points to Identify the Clustering Structure*
PCA	*Principal Component Analysis*
PHOG	*Pyramid Histogram of Oriented Gradients*
RCS	*Reference Color Similarity*

RF	*Random Forest*
ROCK	*Robust Clustering Using Links*
SVM	*Support Vector Machines*
TSNE	*t-distributed Stochastic Neighbor Embedding*

SUMÁRIO

1 INTRODUÇÃO

Sendo o quarto grão mais cultivado do mundo e a principal cultura agrícola brasileira, a soja é um dos alimentos mais importantes no mundo. Em 2018, o Brasil obteve o recorde na produção, com $119{,}3$ milhões de toneladas de soja. Tal valor foi obtido do acréscimo, em relação ao ano de 2017, de $3{,}7\%$ da área cultivada e de 1% na produtividade por hectare (KIST, 2018).

Tais números não seriam alcançados não fossem os avanços científicos e a disponibilização de novas tecnologias ao setor produtivo (FREITAS, 2011). Dentre os avanços destacam-se a mecanização, técnicas de manejo do solo e soluções para prevenção e manejo de pragas e doenças.

Uma das maneiras de obter boas informações para consequentes avanços científicos, são os testes de vigor, utilizados para encontrar diferenças de qualidade entre lotes de sementes durante o armazenamento ou após a semeadura, evidenciando quais foram os melhores e destacando as condições de plantio para os mesmos (SANTANNA, 2014). Um desses testes é o teste de tetrazólio que além de determinar o vigor dos lotes, oferece informações relacionadas às causas de uma redução de qualidade, identificando tipos de danos como: mecânico, deterioração por umidade e percevejos (NETO et al., 2008). Entretanto, a análise para classificar o dano é realizado por um especialista de maneira visual e, considerando a quantidade de amostra em um lote, esse trabalho pode ser inviável. A Figura 1 apresenta exemplos de sementes com diferentes tipos de dano após o teste de tetraozólio.

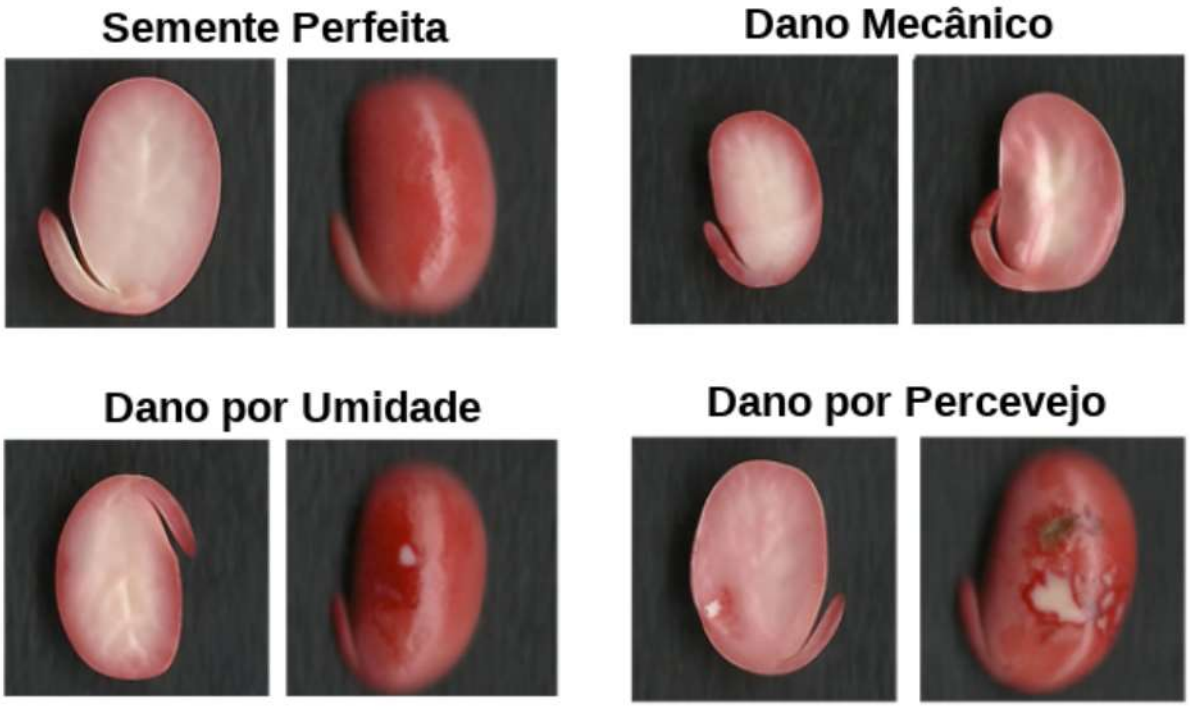
Os seguintes trabalhos (SANTANNA, 2014; PEREIRA, 2019; ALVES, 2018; BRESSAN, 2018; CAMARGO, 2017), evidenciam que técnicas de aprendizado supervisionado, para a classificação do dano após o teste de tetrazólio, podem apresentar bons resultados. Contudo, esses trabalhos não exploram o uso de técnicas não supervisionadas. Desta forma, o presente trabalho tem por objetivo analisar o desempenho de diferentes técnicas de agrupamentos, de forma a possibilitar melhorias na classificação de sementes de soja.

1.1 JUSTIFICATIVA

A análise de dano de sementes de soja submetidas ao teste de tetrazólio é um método que, ao ser realizado manualmente, é um trabalho lento e cansativo, visto que são necessárias algumas horas para ser realizado e é um processo visual (NETO et al., 2008).

Por esta razão e para aumentar a produtividade, trabalhos relacionados (PEREIRA, 2019), (ALVES, 2018), (BRESSAN, 2018), (CAMARGO, 2017) propõem técnicas de aprendizado ativo (classificação supervisionada) para esse tipo de análise. Apesar de apresentarem resultados significativos, os mesmos apresentam como foco técnicas de aprendizado supervisionadas. São pouco exploradas as técnicas de aprendizado não-supervisionadas. As estratégias de

Figura 1 – Exemplos de diferentes tipos de dano visualizados nas sementes após o teste de tetraozólio.



aprendizado ativo que utilizam técnicas de agrupamento como pré-processamento consideram por exemplo o algoritmo k-means. Não é realizada uma avaliação experimental extensiva, comparando os desempenhos de diferentes algoritmos de agrupamento. Portanto, o presente trabalho pretende investigar diferentes técnicas não-supervisionadas, de forma a contribuir também no processo de aprendizado das estratégias de aprendizado ativo.

1.2 OBJETIVOS

Este trabalho tem por objetivo realizar uma avaliação experimental extensiva, considerando diferentes técnicas não-supervisionadas para melhorar o processo de classificação de sementes de soja submetidas ao teste de tetrazólio. Para tanto, os seguintes objetivos específicos devem ser executados:

- Organização dos conjuntos de imagens de sementes de soja;
- Extração e seleção das características que melhor descrevem os conjuntos de imagens;
- Avaliação de desempenho de diferentes técnicas de agrupamento;
- Análise comparativa e validação dos resultados obtidos considerando os diferentes conjuntos de imagens, descritores e agrupamentos.

2 REVISÃO DE LITERATURA

Neste capítulo serão apresentadas as técnicas de descrição e de aprendizado de máquina utilizadas no presente trabalho. Também serao descritors os diferentes métodos utilizados para avaliação de agrupamentos e visualização gráfica dos resultados. Por fim, serão descritos os trabalhos relacionados e como este presente trabalho contribui academicamente para a área.

2.1 APRENDIZADO DE DESCRITORES

Para classificação de um conjunto de imagens é necessário identificar padrões que diferenciem as amostras em classes distintas. Tais padrões podem ser obtidos por meio de diferentes descritores (extratores) de imagens (GONZALES; WOODS, 2017).

Os descritores consideram diferentes propriedades visuais baseados em cor, forma e textura (COSTA; TRAINA, 2012). Sendo assim, cada descritor realiza a extração e gera um vetor de características (valores numéricos), descrevendo as imagens.

2.1.1 EXTRAÇÃO DE CARACTERÍSTICAS

Na literatura existem diversos extratores de características. Os extratores baseados em cor são amplamente utilizados na literatura, especialmente para classificação de imagens naturais. Muitos deles consideram o histograma de cores (SWAIN; BALLARD, 1991), o qual descreve o conteúdo global da imagem de acordo com o percentual de pixels de cada cor.

Alguns exemplos de extratores baseados em cor são: *Auto Color Correlogram* - ACC (HUANG; KUMAR; ZABIH, 1997), *Border/Interior Pixel Classification* - BIC (STEHLING; NASCIMENTO; FALCÃO, 2002), *Color and Edge Directivity Descriptor* - CEDD (CHATZICHRISTOFIS; BOUTALIS, 2008a), *Global Color Histogram* - GCH (STRICKER; OREGON, 1995), *Local Color Histogram* - LCH (SMITH; CHANG, 1996), *Reference Color Similarity* - RCS (SMITH; CHANG, 2011).

Além da característica baseada em cor, uma imagem pode ser representada por meio de dados de textura. Os valores de textura encontrados em uma imagem possuem informações sobre a luminosidade, distribuição espacial e arranjo estrutural da superfície em relação às regiões vizinhas.

Exemplos de extratores baseados em textura são: Gabor (ZHANG et al., 2000), Haralick (HARALICK; SHANMUGAM; DISTEIN, 1973), *Local Binary Pattern* - LBP (GUO; ZHANG; ZHANG, 2010), *Moments* (GRAF, 2015), Medidas de Primeira Ordem - MPO, Medidas de Primeira Ordem de Cor - MPOC, *Pyramid Histogram of Oriented Gradients* - PHOG (BOSCH; ZISSERMAN; MUNOZ, 2007) e Tamura (TAMURA; MORI; YAMAWAKI, 1978).

Alguns extratores combinam diferentes tipos de características. Os extratores *Fuzzy Color and Texture Histogram* - FCTH (CHATZICHRISTOFIS; BOUTALIS, 2008b) e *Join*

Composite Descriptor - JCD (FCTH + CEDD) combinam características baseadas em cor e em textura para descrever as imagens.

2.2 APRENDIZADO DE MÁQUINA

Sendo uma subárea da inteligência artificial, o aprendizado de máquina tem como ênfase o desenvolvimento de sistemas capazes de aprender um determinado padrão ou comportamento automaticamente, a partir de exemplos ou experiência.

Diferentes abordagens de aprendizado supervisionadas e não supervisionadas podem ser consideradas. Tais abordagens são descritas nas seções 2.2.1 e 2.2.2, respectivamente.

2.2.1 CLASSIFICAÇÃO SUPERVISIONADA

Na classificação supervisionada, o processo de aprendizado é realizado por meio de um conjunto de (treinamento de) dados previamente rotulado (CAMPBELL, 2002). Neste processo, um oráculo ou especialista precisa identificar amostras que melhor representam uma classe, para que o algoritmo, após treinado, consiga um melhor resultado (MáXIMO O. A; FERNANDES, 2005).

Existem diversos algoritmos de aprendizado supervisionado propostos na literatura, dentre eles: *Random Forest* - RF (GISLASON; BENEDIKTSSON; SVEINSSON, 2006), *Support Vector Machines* - SVM (VAPNIK; GOLOWICH; SMOLA, 1997) e *Optimum-Path Forest* - OPF (P.; FALCãO; SUZUKI, 2009).

No entanto, no processo de aprendizado supervisionado é importante lidar com alguns desafios, dentre eles: desbalanceamento das amostras em cada uma das classes na base de dados; presença de ruídos na base, como dados imperfeitos ou dados fora da distribuição normal esperada (*outliers*); ajuste excessivo dos dados (*overfitting*) ou generalização insuficiente dos dados (*underfitting*), ausência de valores (*missing values*) para determinadas características. Tais desafios dependem diretamente dos dados utilizados no treinamento do algoritmo de aprendizado.

2.2.2 CLASSIFICAÇÃO NÃO SUPERVISIONADA

Na classificação não supervisionada tem-se o objetivo de encontrar, no espaço de características multidimensional, agrupamentos (*clusters*) de dados, de acordo com determinados critérios de similaridade.

A partir de relações de similaridade, amostras similares entre si são agrupadas em um mesmo *cluster*. Sendo assim, é possível descrever as características inerentes de cada um dos *clusters* encontrados pelo agrupamento, possibilitando um melhor entendimento do conjunto de dados e a classificação de novos dados.

Existem diversos métodos de agrupamentos que podem ser divididos em categorias, tais como: particionais, hierárquicos e baseados em densidade.

Métodos particionais constroem k partições (grupos) de dados. A partir de um particionamento inicial este modelo utiliza a técnica de realocação iterativa para melhorar o particionamento. O principal representante desta categoria é o k-means ou k-médias. Criado por (MACQUEEN, 1967) é um dos mais populares algoritmos de agrupamento. Seu funcionamento pode ser descrito por quatro fases: inicialização; definição dos *clusters* (agrupamentos); movimentação de centroides e otimização. Na inicialização amostras aleatórias são definidas como centroides, que são os pontos centrais dos *cluster*. Na fase de definição dos *clusters* é calculada a distância (e.g. Euclidiana) entre todas as amostras a cada um dos centroides, sendo cada amostra atribuída ao respectivo *cluster* do centroide que apresenta a menor distância calculada. Definidos os *clusters*, na fase de movimentação de centroides é calculada a média das amostras de cada *cluster*. Em seguida, as amostras mais próximas das médias são definidas como novos centroides. Assim, a fase de otimização se resume em executar as fases de definição dos *clusters* e movimentação de centroides repetidas vezes até que os valores centrais dos *clusters* se tornem estáticos, atingindo assim os *clusters* finais.

Similar ao k-means, o algoritmo k-medoids (KAUFMAN; RUSSEEUW, 1987) também tem essa estrutura de fases. Entretanto, em sua fase de movimentação de centroides, não é a amostra mais próxima à média que é definida como centro do *cluster* e sim uma das amostras (denominada medoid) do agrupamento que mais se encontra ao centro. Na otimização a estratégia é realizar tentativas de trocas de medoids e avaliar a qualidade dos novos *clusters*. Em relação ao k-means, este algoritmo apresenta a vantagem de ser menos susceptível ao ruído, uma vez que *outliers* pouco afetam a escolha do medoid. Outros algoritmos particionais são: CLARANS (RAYMOND; HAN, 2002) e FCM (BEZDEK; PAL, 1992).

Métodos hierárquicos são divididos em aglomerativos e divisivos. A ideia deste tipo de método é construir (aglomerativo) ou desconstruir (divisivos) uma árvore binária, em que os nós são amostras do conjunto de dados e suas ligações são realizadas com base na distância (similaridade) dos dados. Exemplos de algoritmos hierárquicos são: AGNES (KAUFMAN; ROUSSEEUW, 1990a), CURE (GUHA; RASTOGI; SHIM, 1998) e ROCK (GUHA; RASTOGI; SHIM, 2000).

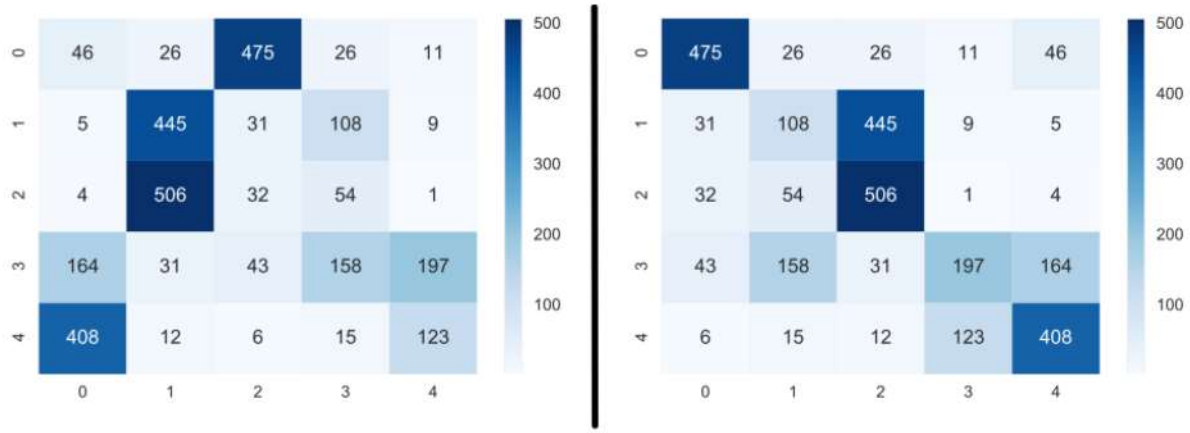
Os métodos baseados em densidade, como o próprio nome diz agrupam de acordo com uma densidade, estabelecida por meio dos parâmetros de entrada que, em geral, ditam qual a densidade mínima necessária para formar um grupo. A principal técnica desta categoria é a *Density Based Spatial Clustering of Application with Noise* - DBSCAN (ESTER et al., 1996). Este algoritmo determina que um grupo possui pontos centrais e pontos de borda. Os centrais são todos aqueles em que dado um raio r possuem o mínimo $MinPts$ de pontos vizinhos e os pontos de borda são os que não possuem o mínimo de vizinhos mas um dos vizinhos é ponto central. A execução desta técnica termina quando nenhum novo ponto pode ser adicionado a um grupo. Outro algoritmo baseado em densidade que é utilizado nesse trabalho é o OPTICS (ANKERSTM; BREUNING, 1999).

2.2.3 AVALIAÇÃO DE AGRUPAMENTOS

Quando se trata de agrupamentos, dois tipos de métricas são consideradas para análise, métodos que se baseiam no rótulo verdadeiro das amostras e os que não utilizam dessa informação. Neste trabalho embora haja conhecimento do rótulo verdadeiro das amostras, serão abordados os dois tipos de métricas para uma análise mais completa. As métricas utilizadas para análise são acurácia, Fowlkes-Mallows (FOWLKES) (FOWLKES; MALLOWS, 1983), Davies-Bouldin (DAVIES) (DAVIES; BOLDIN, 1979) e Calinski-Harabasz (CALINSKI) (CALINSKI; HARABASZ, 1974).

A acurácia consiste no percentual de acerto das amostras. Seu cálculo se dá pela quantidade de amostras agrupadas corretamente dividido pela quantidade total de amostras. As técnicas de agrupamento não se comprometem em definir corretamente os rótulos, somente em agrupar as amostras, ou seja, os grupos formados podem ou não representar adequadamente as diferentes classes do conjunto. Sendo assim, o cálculo de acurácia baseia-se em um método que realiza a organização dos rótulos atribuídos e encontra-se a maior acurácia possível (MORBIEU, 2019). Para isso os dados da matriz de confusão, em que as linhas são os rótulos verdadeiros e as colunas os agrupamentos, são utilizados para encontrar a reorganização das colunas que gera o melhor resultado de acurácia possível, ou seja, uma organização onde os maiores valores de cada linha estão sobre a diagonal principal da matriz. A Figura 2 apresenta exemplos anterior e posterior a uma reorganização da matriz.

Figura 2 – Exemplo de reorganização da matriz para encontrar a melhor acurácia possível de um agrupamento. São apresentados exemplos da matriz, antes (à esquerda) e após (à direita) a organização.



Outra métrica que utiliza o conhecimento dos rótulos verdadeiros é a FOWLKES, a qual é definida como a média geométrica da precisão e da revocação. Sua fórmula é apresentada pela Equação 1 (FOWLKES; MALLOWS, 1983), em que VP corresponde ao número de verdadeiros positivos (amostras de rótulo X que foram agrupadas corretamente), FP consiste no número de falsos positivos (amostras agrupadas com rótulo X mas que não faz parte de X) e FN representa o número de falsos negativos (amostras que deveriam ser agrupadas com rótulo X,

mas não foram). Seus resultados são valores entre 0 a 1, sendo que quanto maior o valor maior à similaridade entre os agrupamentos realizados e as classes verdadeiras.

$$Fowlkes = \frac{VP}{\sqrt{(VP + FP) + (VP + FN)}} \tag{1}$$

Quando não há o conhecimento das classes verdadeiras das amostras são utilizadas métricas que avaliam a qualidade dos agrupamentos formados. Para isso, medidas de coesão e separação são utilizadas, em que coesão refere-se à distância entre as amostras de um mesmo grupo e separação indica a distância de um grupo a outro (RENDóN, 2011).

A métrica DAVIES indica quão boa é a separação entre agrupamentos. Para isso, compara-se a distância entre grupos com o tamanho dos mesmos. Sua fórmula é apresentada pela Equação 2 (DAVIES; BOLDIN, 1979), em que k representa a quantidade total de grupos formados e R_{ij} a comparação entre dois grupos, apresentada pela Equação 3. Onde s_i representa a média da distância de cada ponto do grupo i ao seu *centroid* e d_{ij} representa a distância entre os *centroids* dos grupos i e j. Seus resultados vão de 0 a 1, sendo que menores valores indicam agrupamentos melhores.

$$Davies = \frac{1}{k} \sum_{i=1}^{k} \underset{i \neq j}{max} R_{ij} \tag{2}$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{3}$$

CALINSKI é outra métrica que não utiliza o conhecimento dos rótulos verdadeiros. Nesse caso, valores elevados dessa métrica indicam melhores definições dos agrupamentos, ou seja, grupos densos e bem separados. Seu valor é gerado pela razão do somatório de dispersão entre grupos e a dispersão interna de cada grupo (CALINSKI; HARABASZ, 1974).

Para um *dataset* E de tamanho n_E com k agrupamentos realizados, o valor de CALINSKI é apresentado pela Equação 4 (CALINSKI; HARABASZ, 1974), em que $tr(B_k)$ representa o traço da matriz de dispersão entre grupos e $tr(W_k)$ consiste no traço da matriz de dispersão interna dos grupos, definidos pelas Equações 5 e 6. Nesse caso, C_i representa o conjunto de amostras do grupo i, c_i o centro do grupo i, c_E o centro de E e n_i a quantidade de amostras no grupo i.

$$Calinski = \frac{tr(B_k)}{tr(W_k)} * \frac{n_E - k}{k - 1} \tag{4}$$

$$W_k = \sum_{i=1}^{k} \sum_{x \in C_i} (x - c_i)(x - c_i)^T \tag{5}$$

$$B_k = \sum_{i=1}^{k} n_i (c_i - c_E)(c_i - c_E)^T \tag{6}$$

2.2.4 REDUÇÃO DE DIMENSIONALIDADE

Os descritores apresentados na Seção 2.1.1 caracterizam as imagens em vetores de múltiplas dimensões, dessa maneira impossibilita a visualização gráfica das amostras com os valores reais. A fim de resolver esse problema métodos de redução de dimensionalidade são aplicados para reduzir as amostras em duas ou três dimensões (PERPINAN, 2001).

As técnicas de redução de dimensionalidade (ou redução de características) são divididas em duas categorias, métodos de seleção de características e os de extração de características (PALIWAL, 1992). Os métodos de seleção baseiam-se em avaliar quais características da amostra são as mais importantes para que possam representá-la. Os métodos de extração utilizam todas as características da amostra para calcular uma nova representação e são mais indicados para realizar a representação gráfica.

Dentre as técnicas de redução de características baseadas na extração, há os que utilizam funções lineares e não lineares. *Principal Component Analysis* (PCA) (TIPPING; BISHOP, 1999) é uma técnica que executa um mapeamento linear dos dados para redução de dimensão, de forma que a variação dos dados, na representação com menor dimensão, seja maximizada. *T-distributed Stochastic Neighbor Embedding* (TSNE) (MAATEN; HINTON, 2008) é uma técnica não linear que calcula a probabilidade das similaridades das características e procura minimizar a divergência entre as probabilidades da amostra reduzida e a original.

2.3 TRABALHOS RELACIONADOS

Alguns trabalhos têm sido desenvolvidos para classificação de sementes de soja submetidas ao teste de tetrazólio.

Em (PEREIRA, 2019), é apresentada uma proposta de classificação por meio de estratégias de aprendizado ativo. Neste processo é utilizada a técnica de agrupamento apenas para separar amostras a serem rotuladas pelo especialista, que sejam mais informativas ao aprendizado do classificador supervisionado. O foco do trabalho consistiu no aprendizado supervisionado. O trabalho originou alguns subconjuntos de imagens de sementes de soja, os quais serão considerados no presente trabalho.

De forma similar, em (BRESSAN, 2018), (CAMARGO, 2017) e (ALVES, 2018), uma técnica de agrupamento também é considerada para seleção de amostras mais informativas pelas estratégias de aprendizado ativo propostas. Os trabalhos exploram apenas a técnica k-means e apenas como um pré-processamento. Tais trabalhos têm como foco o aprendizado supervisionado. Mais especificamente, (BRESSAN, 2018) desenvolveu abordagens para recuperação baseada em conteúdo e (CAMARGO, 2017) propôs abordagens semi-supervisionadas. Já (ALVES, 2018) explorou abordagens de aprendizado profundo para geração de imagens sintéticas.

Apesar de apresentarem resultados significativos, os trabalhos relacionados citados acima apresentam como foco técnicas de aprendizado supervisionadas. São pouco exploradas as técnicas de aprendizado não supervisionadas. Tais trabalhos se limitam a utilizar apenas a

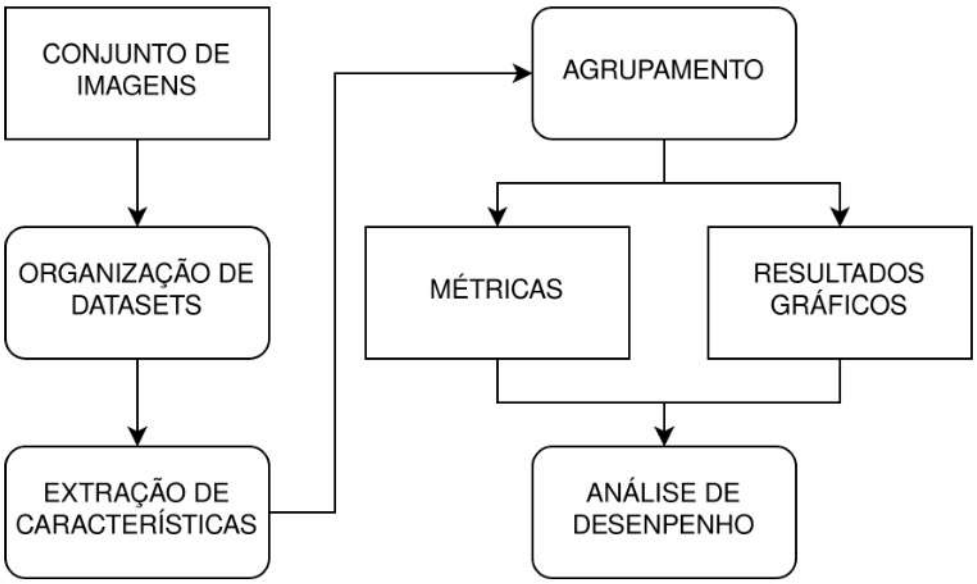
técnica k-means. Portanto, o presente trabalho apresenta como contribuição a realização de uma avaliação experimental extensiva, comparando e validando o desempenho de diferentes algoritmos de agrupamento em diferentes subconjuntos de sementes de soja. Dessa forma, é possível investigar as estratégias de agrupamento mais adequadas para cada subconjunto. É possível também melhorar os processos de aprendizado das estratégias de aprendizado ativo, pois quão melhor for o agrupamento realizado em suas etapas intermediárias, melhor será o resultado final de classificação de imagens.

3 METODOLOGIA

Neste capítulo é apresentada a metodologia, descrevendo o passo a passo de cada processo realizado. Serão apresentadas a descrição dos subconjuntos de imagens de sementes de soja (Seção 3.1), a descrição dos cenários (Seção 3.2), incluindo os descritores e configurações dos algoritmos a serem considerados no trabalho, bem como os resultados adquiridos (Seção 3.3).

Para alcançar o objetivo de analisar técnicas de agrupamento para classificação de imagens de sementes de soja, são utilizados os algoritmos de descrição mencionados na Seção 2.1.1, juntamente com as diferentes técnicas de agrupamento descritas na Seção 2.2.2. A Figura 3 apresenta o pipeline da metodologia.

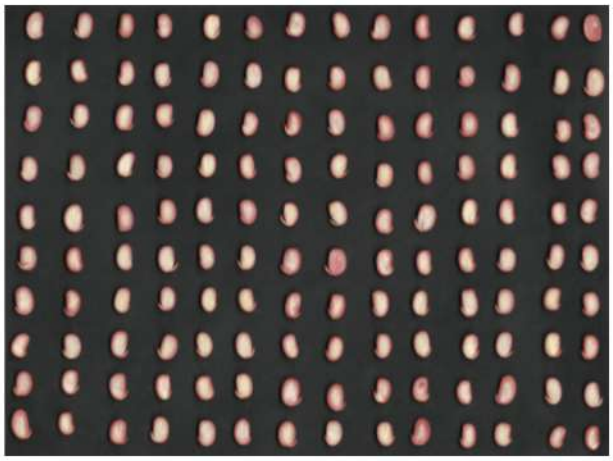
Figura 3 – Pipeline da metodologia.



A fim de contribuir com os trabalhos relacionados, os conjuntos de imagens utilizados são os mesmos apresentados em (PEREIRA, 2019). Em (PEREIRA, 2019) é constatado que após o teste de tetrazólio, realizado por especialistas na unidade da empresa Belágricola em Tamarana-PR, foi utilizado o scanner EPSON L355 na resolução de 1200dpi para adquirir as imagens das sementes dispostas em matriz sobre um fundo preto, conforme mostra a figura 4. Para finalizar a obtenção dos dados, com o intuito de criar um conjunto de imagens individuais das sementes, as imagens obtidas passam por um pré-processamento que pode ser dividido em: segmentação, identificação e recorte (PEREIRA, 2019).

As imagens das sementes podem ser organizadas de diversas maneiras, resultando em diferentes *datasets*, para avaliação experimental. Os detalhes sobre os conjuntos construídos e considerados no presente projeto são apresentados em 3.1. Nesta etapa também foram retiradas do conjunto imagens com mancha e problemas de recorte.

Figura 4 – Exemplos de uma lâmina com imagens de sementes de soja submetidas ao teste de tetrazólio.



Fonte: (PEREIRA, 2019)

Após a organização dos subconjuntos, realiza-se a extração das características. Nesta fase, cada imagem de semente de soja é representada por um vetor de características, considerando cada descritor apresentado na Seção 2.1.

Para cada extração de característica realizada, são executadas as técnicas de agrupamento mencionadas na Seção 2.2.2. Como resultado tem-se as métricas mencionadas na Seção 2.2.3 e gráficos dos agrupamentos realizados. A partir disso, é possível analisar o desempenho dos algoritmos e descritores na classificação de sementes de soja.

3.1 DESCRIÇÃO DOS CONJUNTOS DE IMAGENS

Os analistas, após o teste de tetrazólio, passam por um processo visual em que identificam e rotulam cada porção de uma semente de soja (a qual contém duas partes internas e duas partes externas) em classes. A classe é definida pelo tipo de dano e o nível de gravidade do mesmo.

Em (PEREIRA, 2019) duas aquisições de imagens foram realizadas na empresa Belagricola. Na primeira foram obtidas 1.400 imagens e na segunda 1.358 (cada imagem referente a uma das 4 porções de uma semente). A partir dessas duas aquisições, diferentes subconjuntos foram criados na medida em que os especialistas rotularam as imagens.

Este trabalho considera os conjuntos usados em (PEREIRA, 2019) em que cada porção da semente é rotulada separadamente, dado que foi observado um melhor desempenho para esse cenário. Os conjuntos foram identificados neste trabalho como D_6, D_7 e D_8 para estabelecer uma relação com os apresentados em (PEREIRA, 2019). Nesse trabalho foram criados os conjuntos D_9 e D_{10} que são uma reorganização das classes do conjunto D_8. As Tabelas 1-5 apresentam a descrição e a quantidade de amostras em cada classe dos conjuntos.

Os conjuntos D_6 e D_7 tratam-se da primeira e segunda aquisição respectivamente,

Tabela 1 – Descrições das classes e distribuição das amostras em cada classe referentes ao conjunto D_6.

Classes	Descrição	Amostras
OXE	Porção Externa Perfeita	483
OXI	Porção Interna Perfeita	499
2UE	Porção Externa com Dano por Umidade nível 2	23
2UI	Porção Interna com Dano por Umidade nível 2	7
3ME	Porção Externa com Dano Mecânico nível 3	32
3MI	Porção Interna com Dano Mecânico nível 3	28
3PE	Porção Externa com Dano por Percevejo nível 3	78
3PI	Porção Interna com Dano por Percevejo nível 3	36
3UE	Porção Externa com Dano por Umidade nível 3	36
3UI	Porção Interna com Dano por Umidade nível 3	47

Tabela 2 – Descrições das classes e distribuição das amostras em cada classe referentes ao conjunto D_7.

Classes	Descrição	Amostras
OXE	Porção Externa Perfeita	306
OXI	Porção Interna Perfeita	374
3ME	Porção Externa com Dano Mecânico nível 3	4
3MI	Porção Interna com Dano Mecânico nível 3	5
3PE	Porção Externa com Dano por Percevejo nível 3	18
3PI	Porção Interna com Dano por Percevejo nível 3	17
3UI	Porção Interna com Dano por Umidade nível 3	9

Tabela 3 – Descrições das classes e distribuição das amostras em cada classe referentes ao conjunto D_8.

Classes	Descrição	Amostras
OXE	Porção Externa Perfeita	789
OXI	Porção Interna Perfeita	873
2UE	Porção Externa com Dano por Umidade nível 2	23
2UI	Porção Interna com Dano por Umidade nível 2	7
3ME	Porção Externa com Dano Mecânico nível 3	36
3MI	Porção Interna com Dano Mecânico nível 3	33
3PE	Porção Externa com Dano por Percevejo nível 3	96
3PI	Porção Interna com Dano por Percevejo nível 3	53
3UE	Porção Externa com Dano por Umidade nível 3	36
3UI	Porção Interna com Dano por Umidade nível 3	56

sendo selecionadas apenas as amostras com classes de nível de dano entre 0 e 3. No entanto, para o presente projeto considerando o conjunto D_6, na etapa de organização de *datasets*, foram retiradas imagens que apresentam manchas e problemas de recorte, visto que podem impactar no aprendizado e desempenho dos algoritmos.

O conjunto D_8 é a junção dos conjuntos D_6 e D_7. Visto que em D_8 encontra-se todas as imagens utilizadas, a partir dele foram criados os conjuntos D_9 e D_{10}. D_9 é uma

Tabela 4 – Descrições das classes e distribuição das amostras em cada classe referentes ao conjunto D_9.

Classes	Descrição	Amostras
XE	Porção Externa Perfeita	789
XI	Porção Interna Perfeita	873
UE	Porção Externa com Dano por Umidade	59
UI	Porção Interna com Dano por Umidade	63
ME	Porção Externa com Dano Mecânico	36
MI	Porção Interna com Dano Mecânico	33
PE	Porção Externa com Dano por Percevejo	96
PI	Porção Interna com Dano por Percevejo	53

Tabela 5 – Descrições das classes e distribuição das amostras em cada classe referentes ao conjunto D_{10}

Classes	Descrição	Amostras
X	Porção Perfeita	1662
U	Porção com Dano por Umidade	122
M	Porção com Dano Mecânico	69
P	Porção com Dano por Percevejo	149

Tabela 6 – Quantidades de classes e amostras para os conjuntos utilizados.

Conjuntos de Imagens	Classes	Amostras
D_6	10	1269
D_7	7	733
D_8	10	2002
D_9	8	2002
D_{10}	4	2002

reorganização das classes em que retira-se a existência de níveis do conjunto D_8 (i.e. são considerados apenas os tipos de danos), enquanto D_{10} retira-se os níveis e a diferenciação de porção interna e externa.

As criações de novos *datasets* foram realizadas para mensurar como a alteração nas quantidades de amostras e de classes impactam nos experimentos.

3.2 DESCRIÇÃO DOS CENÁRIOS

Para realizar os experimentos, alguns cenários são definidos. A descrição de cada conjunto de dados foi realizada por meio de um programa desenvolvido por (BRESSAN, 2018). A partir de imagens no formato do tipo .png ou .tiff como entrada, é possível obter a descrição dos conjuntos de dados mencionados na Seção 3.1. A Tabela 7 apresenta todos os descritores utilizados.

Após a descrição dos conjuntos, os algoritmos de agrupamento foram executados para análises e coletas das métricas e gráficos. Este processo foi realizado por meio da biblioteca de

Tabela 7 – Descritores utilizados para extração das características das imagens de sementes de soja.

Técnica	Descrição	Categoria	Características
ACC	*Auto Color Correlogram*	Cor	768
BIC	*Border/Interior Classification*	Cor	128
CEDD	*Color and Edge Directivity Descriptor*	Cor	144
FCTH	*Fuzzy Color and Texture Histogram*	Cor e Textura	192
GABOR	*Gabor Texture Features*	Textura	60
GCH	*Global Color Histogram*	Cor	255
HARACOLOR	*Haralick Color*	Cor	14
HARAFULL	*Haralick Full*	Cor e Textura	14
HARALICK	*Haralick*	Textura	14
JCD	*Join Composite Descriptor*	Cor e Textura	336
LBP	*Local Binary Patterns*	Textura	256
LCH	*Local Color Histogram*	Cor	135
MOMENTS	*Moments*	Textura	4
MPO	Medidas de Primeira Ordem	Textura	6
MPOC	Medidas de Primeira Ordem - Color	Textura	18
PHOG	*Pyramid Histogram of Oriented Gradients*	Textura	40
RCS	*Reference Color Similarity*	Cor	77
TAMURA	*Tamura Descriptor*	Textura	18

código aberto de mineração de dados *pyclustering* (NOVIKOV, 2019), utilizando a linguagem de programação Python. A Tabela 8 apresenta as técnicas de agrupamento analisadas.

Tabela 8 – Algoritmos de agrupamento utilizados na avaliação experimental.

Técnica	Descrição
AGNES	*Agglomerative Nesting*
CLARANS	*Clustering Large Applications based on Randomized Search*
CURE	*Clustering using Representatives*
DBSCAN	*Density Based Spatial Clustering of Application with Noise*
FCM	*Fuzzy-c-Means*
K-Means	*K-Means*
K-Medoid	*K-Medoid*
OPTICS	*Ordering Points to Identify the Clustering Structure*
ROCK	*Robust Clustering Using Links*

Para os parâmetros de entrada das técnicas de agrupamento, foram utilizados valores *default*, ou seja, o padrão da literatura. Há algoritmos que necessitam do parâmetro referente ao *número de grupos*. Nesses casos, foi utilizado o número de classes existentes nos conjuntos.

Como visto na Seção 2.2.2, algoritmos de densidade necessitam do parâmetro *raio* e *mínimo de pontos vizinhos*. Para definir o valor do *raio* foi utilizada a técnica apresentada em (ESTER et al., 1996), enquanto que para definir o *mínimo de pontos vizinhos*, que é um número inteiro, foram realizadas múltiplas experimentações, alterando o valor do parâmetro. Em seguida, foi escolhido o valor em que o número de agrupamentos resultante fosse o mais

próximo do total de classes.

Com exceção da acurácia, foi utilizada a biblioteca *sklearn* para coletar as métricas apresentadas em 2.2.3. Para definir a acurácia foi realizado um método de organização dos rótulos agrupados (MORBIEU, 2019), como visto na seção 2.2.3.

3.3 RESULTADOS

A partir dos resultados de acurácia, é possível realizar as primeiras análises de desempenho. As Tabelas 9-13 apresentam os resultados de acurácias obtidos por cada descritor e algoritmo de agrupamento em cada conjunto de dados (D_6-D_{10}, respectivamente). Em negrito destaca-se, para cada descritor, qual algoritmo de agrupamento obteve o melhor resultado, enquanto que sublinhado destaca-se, para cada algoritmo de agrupamento, qual descritor apresentou um melhor desempenho. O valor com asterisco representa a melhor combinação (par descritor e algoritmo de agrupamento) em que foi obtido um maior valor de acurácia, considerando todos os descritores e classificadores.

Tabela 9 – Resultados de acurácias obtidas por cada descritor e algoritmo de agrupamento considerando o conjunto D_6. Os melhores resultados (i.e. algoritmos de agrupamento) para cada descritor são destacados em negrito. O melhor resultado (i.e. descritores) para cada algoritmo de agrupamento são apresentados sublinhados. O melhor resultado (i.e maior acurácia) obtido é apresentado com asterisco. São apresentadas também as médias de acurácias obtidas considerando todos os descritores e classificadores.

–	AGNES	CLARANS	CURE	DBSCAN	FCM	KMEANS	KMEDOIDS	OPTICS	ROCK	MÉDIA	D.PADRÃO
ACC	0,578	0,286	0,711	0,704	0,309	0,336	0,491	0,704	**0,745**	0,540	0,190
BIC	0,704	0,370	**0,744**	0,716	0,271	0,318	0,502	0,716	0,388	0,525	0,195
CEDD	0,630	0,337	**0,737**	0,390	0,423	0,331	0,664	0,390	0,394	0,447	0,155
FCTH	0,723	0,422	**0,733**	0,537	0,350	0,451	0,644	0,537	0,723	0,569	0,144
GABOR	0,404	0,283	0,473	0,394	0,260	0,281	0,289	0,394	**0,552**	0,370	0,100
GCH	0,697	0,405	***0,768**	0,714	0,323	0,316	0,537	0,714	0,386	0,540	0,186
HARACOLOR	0,582	0,377	0,652	0,703	0,322	0,360	0,383	0,703	**0,722**	0,534	0,140
HARAFULL	**0,544**	0,258	0,534	0,391	0,297	0,319	0,298	0,398	0,390	0,381	0,170
HARALICK	0,471	0,293	0,589	0,385	0,350	0,366	0,355	0,393	**0,733**	0,437	0,102
JCD	0,721	0,316	**0,733**	0,392	0,395	0,389	0,650	0,392	0,395	0,487	0,164
LBP	0,277	0,195	0,341	0,391	0,169	0,191	0,198	0,391	**0,397**	0,283	0,098
LCH	0,567	0,366	0,392	0,583	0,407	0,341	**0,603**	0,583	0,394	0,471	0,109
MOMENTS	0,630	0,429	**0,724**	0,723	0,267	0,319	0,362	0,723	**0,724**	0,545	0,197
MPO	0,608	0,362	0,600	0,553	0,316	0,330	0,326	0,553	**0,685**	0,482	0,146
MPOC	0,597	0,287	**0,717**	0,389	0,261	0,300	0,301	0,389	0,407	0,405	0,155
PHOG	0,437	0,243	0,396	0,394	0,363	0,278	**0,528**	0,394	0,396	0,381	0,083
RCS	0,593	0,371	0,723	0,701	0,268	0,354	0,477	0,701	**0,746**	0,548	0,184
TAMURA	**0,536**	0,275	0,392	0,394	0,285	0,284	0,496	0,394	0,390	0,383	0,092
MÉDIA	0,572	0,326	0,609	0,525	0,313	0,326	0,450	0,526	0,531	-	

Nota-se que, para a maioria dos conjuntos de dados e descritores, os algoritmos CURE e ROCK se destacam, apresentando os melhores resultados. Vale ressaltar que AGNES também se destacou para alguns conjuntos. No entanto, AGNES apresentou valores de acurácias menores em relação aos apresentados por CURE e ROCK.

Com relação aos descritores, nenhum se destacou simultaneamente entre a maioria dos algoritmos de agrupamento. No entanto, de forma geral, os descritores FCTH, GCH e RCS

Tabela 10 – Resultados de acurácias obtidas por cada descritor e algoritmo de agrupamento considerando o conjunto D_7. Os melhores resultados (i.e. algoritmos de agrupamento) para cada descritor são destacados em negrito. Os melhores resultados (i.e. descritores) para cada algoritmo de agrupamento são apresentados sublinhados. O melhor resultado (i.e maior acurácia) obtido é apresentado com asterisco. São apresentadas também as médias de acurácias obtidas considerando todos os descritores e classificadores.

–	AGNES	CLARANS	CURE	DBSCAN	FCM	KMEANS	KMEDOIDS	OPTICS	ROCK	MÉDIA	D.PADRÃO
ACC	0,776	0,523	**0,853**	0,842	0,398	0,517	0,437	0,840	0,502	0,632	0,191
BIC	0,681	0,467	**0,920**	0,881	0,409	0,458	0,608	0,880	0,509	0,646	0,203
CEDD	0,753	0,516	**0,920**	0,790	0,444	0,466	0,783	0,854	0,512	0,671	0,184
FCTH	0,653	0,592	***0,926**	0,508	0,541	0,571	0,606	0,508	0,502	0,601	0,133
GABOR	0,658	0,400	0,696	**0,853**	0,416	0,430	0,477	0,853	0,759	0,616	0,188
GCH	0,778	0,468	**0,920**	0,873	0,395	0,386	0,745	0,873	0,505	0,660	0,220
HARACOLOR	0,588	0,367	0,675	0,844	0,373	0,405	0,412	0,859	**0,888**	0,601	0,222
HARAFULL	0,630	0,456	0,628	**0,636**	0,392	0,389	0,357	0,636	0,613	0,526	0,124
HARALICK	0,573	0,427	0,630	0,465	0,420	0,435	0,426	0,521	**0,905**	0,533	0,158
JCD	0,750	0,409	**0,917**	0,866	0,405	0,515	0,626	0,859	0,506	0,651	0,203
LBP	0,416	0,237	0,411	0,505	0,239	0,257	0,286	0,492	**0,506**	0,372	0,117
LCH	0,741	0,472	0,513	0,527	0,469	0,414	**0,742**	0,580	0,012	0,497	0,215
MOMENTS	0,614	0,355	0,675	**0,844**	0,385	0,409	0,420	0,844	0,817	0,596	0,208
MPO	0,636	0,416	0,584	0,546	0,426	0,425	0,383	**0,681**	0,643	0,527	0,115
MPOC	0,540	0,411	**0,802**	0,523	0,344	0,381	0,435	0,520	0,538	0,499	0,135
PHOG	**0,602**	0,398	0,509	0,516	0,371	0,339	0,416	0,572	0,509	0,470	0,092
RCS	0,576	0,412	**0,917**	0,842	0,351	0,445	0,408	0,865	**0,917**	0,637	0,244
TAMURA	**0,559**	0,368	0,512	0,510	0,349	0,392	0,442	0,510	0,510	0,461	0,076
MÉDIA	0,640	0,427	0,723	0,687	0,396	0,424	0,501	0,708	0,592	-	

atingiram os melhores valores de acurácia, principalmente quando utilizados com os algoritmos AGNES, CURE e ROCK.

Como no conjunto D_{10} as amostras internas e externas são consideradas como sendo de um mesmo grupo, observa-se um comportamento diferente em relação ao apresentado em outros conjuntos. Por exemplo, os algoritmos OPTICS e ROCK se destacam, apresentando maiores acurácias para uma quantidade maior de descritores. No entanto, ROCK apresenta o melhor resultado, atingindo acurácias de até $83{,}1\%$.

Além das acurácias, foram realizadas análises de desempenho utilizando outras métricas (FOWLKES, DAVIES e CALINSKI). As Tabelas 14-18 apresentam os resultados obtidos para cada descritor, considerando os conjuntos D_6 e D_7 com CURE e os conjuntos D_8, D_9 e D_{10} com ROCK, visto que essas foram as melhores combinações (par descritor e algoritmo de agrupamento). Em negrito estão destacados os melhores resultados (descritores) obtidos de acordo com cada métrica. Os apêndices A-E apresentam os resultados obtidos, incluindo a quantidade de agrupamentos gerados e os valores das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) para cada descritor, considerando todos os algoritmos de agrupamento e conjuntos de dados.

Pode-se observar que a métrica FOWLKES apresenta resultados similares à acurácia. Observa-se também que na maioria dos casos quando as métricas DAVIES e CALINSKI atingem seus melhores resultados (valores destacados em negrito), os valores de acurácia são inferiores à melhor acurácia obtida para o agrupamento. Isso pode indicar que métricas baseadas em coesão e separação de agrupamentos, neste contexto, não devem ser utilizadas para analisar a

Tabela 11 – Resultados de acurácias obtidas por cada descritor e algoritmo de agrupamento considerando o conjunto D_8. Os melhores resultados (i.e. algoritmos de agrupamento) para cada descritor são destacados em negrito. Os melhores resultados (i.e. descritores) para cada algoritmo de agrupamento são apresentados sublinhados. O melhor resultado (i.e maior acurácia) obtido é apresentado com asterisco. São apresentadas também as médias de acurácias obtidas considerando todos os descritores e classificadores.

–	AGNES	CLARANS	CURE	DBSCAN	FCM	KMEANS	KMEDOIDS	OPTICS	ROCK	MÉDIA	D.PADRÃO
ACC	0,549	0,324	0,654	0,575	0,312	0,340	0,582	0,576	**0,784**	0,522	0,163
BIC	**0,504**	0,411	0,486	0,429	0,280	0,313	0,409	0,429	0,434	0,410	0,073
CEDD	**0,740**	0,315	0,436	0,433	0,364	0,326	0,476	0,433	0,437	0,440	0,125
FCTH	0,682	0,392	**0,785**	0,431	0,335	0,404	0,477	0,431	0,432	0,485	0,148
GABOR	0,325	0,297	0,322	**0,487**	0,219	0,234	0,232	0,487	0,434	0,337	0,107
GCH	0,624	0,324	0,639	**0,640**	0,315	0,338	0,437	0,640	0,437	0,488	0,147
HARACOLOR	0,431	0,342	**0,475**	0,424	0,280	0,306	0,355	0,424	0,463	0,389	0,070
HARAFULL	0,520	0,346	0,402	**0,520**	0,290	0,304	0,295	0,520	0,426	0,402	0,099
HARALICK	0,364	0,285	0,334	0,458	0,289	0,283	0,277	0,457	**0,463**	0,357	0,082
JCD	**0,679**	0,303	0,432	0,434	0,345	0,361	0,415	0,434	0,434	0,426	0,106
LBP	0,257	0,185	0,381	**0,436**	0,162	0,173	0,171	0,436	0,434	0,293	0,126
LCH	**0,626**	0,342	0,437	0,433	0,511	0,371	0,624	0,433	0,437	0,468	0,101
MOMENTS	0,454	0,285	**0,471**	0,459	0,250	0,271	0,270	0,460	0,457	0,375	0,101
MPO	0,408	0,271	0,363	0,466	0,275	0,273	0,260	**0,468**	0,453	0,360	0,091
MPOC	0,393	0,294	**0,463**	0,420	0,257	0,276	0,354	0,421	0,448	0,369	0,077
PHOG	**0,478**	0,301	0,436	0,437	0,374	0,261	0,450	0,437	0,437	0,401	0,074
RCS	0,488	0,334	0,810	0,470	0,273	0,320	0,449	0,470	***0,811**	0,492	0,196
TAMURA	**0,549**	0,259	0,435	0,436	0,284	0,271	0,539	0,436	0,435	0,405	0,110
MÉDIA	0,504	0,312	0,487	0,466	0,301	0,301	0,393	0,466	0,481		

acertabilidade.

Para melhor entendimento dos resultados apresentados pelas métricas é importante observar a representação gráfica dos agrupamentos obtidos. Para tanto, foram consideradas as técnicas de redução de dimensionalidade PCA e TSNE (Figuras 5 e 6), respectivamente. As Figuras 5-6 apresentam os resultados dos agrupamentos destacados na Tabela 17 (i.e. as melhores combinações – par descritor e algoritmo de agrupamento – de acordo com cada métrica). As cores representam os agrupamentos obtidos, enquanto que os símbolos representam as classes verdadeiras das amostras.

Como visto na Tabela 4, a maioria (83%) das amostras do conjunto são da classe perfeita, externa e interna. É possível observar os agrupamentos das amostras dessas classes, representadas pelos símbolos círculo e losango (Figuras 5 e 6). Ambas as técnicas PCA e TSNE não conseguiram agrupar adequadamente as demais classes de danos. No agrupamento apresentado pela Figura 5(a), mesmo com 81% de acurácia, nota-se que a maioria das classes de danos por umidade, percevejo e mecânico, são consideradas perfeitas, ou seja, quase todos os acertos obtidos referem-se às classes perfeitas.

Com relação à métrica DAVIES, pode-se concluir a partir da análise da Figura 5(b), que os resultados que se destacam para tal métrica (e.g. valor de $0{,}35$ para a combinação TAMURA e ROCK na Tabela 17) não indicam um bom agrupamento, visto que formou-se um grupo com quase todas as amostras. Nesse caso, de fato, a acurácia apresentada pelo agrupamento obtido foi de 44%. Sendo assim, DAVIES não é uma boa métrica para se utilizar nesse contexto.

Tabela 12 – Resultados de acurácias obtidas por cada descritor e algoritmo de agrupamento considerando o conjunto D_9. Os melhores resultados (i.e. algoritmos de agrupamento) para cada descritor são destacados em negrito. Os melhores resultados (i.e. descritores) para cada algoritmo de agrupamento são apresentados sublinhados. O melhor resultado (i.e maior acurácia) obtido é apresentado com asterisco. São apresentadas também as médias de acurácias obtidas considerando todos os descritores e classificadores.

–	AGNES	CLARANS	CURE	DBSCAN	FCM	KMEANS	KMEDOIDS	OPTICS	ROCK	MÉDIA	D.PADRÃO
ACC	0,549	0,391	**0,798**	0,575	0,371	0,374	0,587	0,576	0,784	0,556	0,161
BIC	**0,503**	0,366	0,478	0,429	0,328	0,370	0,418	0,429	0,435	0,417	0,055
CEDD	**0,740**	0,347	0,436	0,433	0,366	0,369	0,476	0,433	0,437	0,448	0,117
FCTH	0,682	0,402	**0,784**	0,431	0,370	0,470	0,483	0,431	0,433	0,498	0,139
GABOR	0,352	0,287	0,379	0,495	0,250	0,265	0,251	**0,503**	0,441	0,358	0,103
GCH	0,623	0,357	0,639	**0,640**	0,334	0,415	0,437	0,640	0,437	0,502	0,131
HARACOLOR	**0,504**	0,385	0,476	0,425	0,328	0,368	0,409	0,425	0,469	0,421	0,056
HARAFULL	**0,597**	0,368	0,532	0,487	0,327	0,349	0,320	0,486	0,426	0,432	0,099
HARALICK	0,465	0,356	0,381	0,454	0,309	0,326	0,319	0,454	**0,471**	0,393	0,068
JCD	**0,684**	0,524	0,431	0,434	0,350	0,402	0,415	0,434	0,434	0,456	0,097
LBP	0,407	0,229	0,380	**0,436**	0,185	0,201	0,194	0,436	0,434	0,322	0,116
LCH	**0,633**	0,393	0,437	0,433	0,521	0,390	0,624	0,433	0,437	0,478	0,093
MOMENTS	**0,487**	0,401	0,472	0,406	0,316	0,319	0,336	0,419	0,474	0,403	0,067
MPO	0,441	0,320	0,417	0,466	0,304	0,308	0,302	0,466	**0,482**	0,389	0,079
MPOC	0,396	0,261	0,424	0,423	0,299	0,322	0,315	0,423	**0,455**	0,369	0,070
PHOG	**0,528**	0,325	0,436	0,437	0,391	0,291	0,451	0,437	0,437	0,415	0,071
RCS	0,491	0,462	0,810	0,468	0,329	0,366	0,455	0,468	***0,811**	0,518	0,174
TAMURA	**0,549**	0,395	0,436	0,436	0,314	0,316	0,539	0,436	0,436	0,428	0,082
MÉDIA	0,535	0,365	0,508	0,461	0,333	0,346	0,407	0,463	0,485	-	

A partir da análise da Figura 5(c), pode-se concluir que os resultados que se destacam para métrica CALINSKI (e.g. valor de $6614{,}24$ para combinação MPO e ROCK na Tabela 17) não indicam um bom agrupamento, visto que a maioria das amostras das duas classes perfeitas (XE e XI, ilustradas pelos círculo e losango, respectivamente), as quais representam (83%) das amostras, foram divididas entre cinco grupos, impactando diretamente na acurácia (48%). Desta maneira, CALINSKI também não é uma boa métrica para se utilizar nesse contexto.

Tabela 13 – Resultados de acurácias obtidas por cada descritor e algoritmo de agrupamento considerando o conjunto D_{10}. Os melhores resultados (i.e. algoritmos de agrupamento) para cada descritor são destacados em negrito. Os melhores resultados (i.e. descritores) para cada algoritmo de agrupamento são apresentados sublinhados. O melhor resultado (i.e maior acurácia) obtido é apresentado com asterisco. São apresentadas também as médias de acurácias obtidas considerando todos os descritores e classificadores.

–	AGNES	CLARANS	CURE	DBSCAN	FCM	KMEANS	KMEDOIDS	OPTICS	ROCK	MÉDIA	D.PADRÃO
ACC	0,446	0,294	0,463	0,685	0,297	0,319	0,448	**0,686**	0,462	0,455	0,148
BIC	0,448	0,314	0,829	0,829	0,308	0,331	0,444	0,829	**0,830**	0,574	0,248
CEDD	0,506	0,316	**0,831**	0,829	0,289	0,384	0,315	0,829	0,829	0,570	0,254
FCTH	0,557	0,355	0,468	0,827	0,444	0,448	0,318	0,828	**0,829**	0,564	0,209
GABOR	0,391	0,288	0,451	0,806	0,292	0,302	0,301	**0,809**	0,673	0,479	0,223
GCH	0,783	0,300	0,612	0,445	0,291	0,322	0,420	0,445	**0,830**	0,494	0,203
HARACOLOR	0,329	0,304	0,295	0,661	0,302	0,301	0,301	**0,673**	0,303	0,385	0,160
HARAFULL	0,605	0,425	0,480	0,804	0,387	0,429	0,359	0,814	**0,820**	0,569	0,195
HARALICK	0,485	0,328	0,442	0,800	0,314	0,310	0,310	**0,807**	0,636	0,492	0,207
JCD	0,505	0,433	0,829	**0,830**	0,296	0,416	0,356	0,830	0,829	0,592	0,233
LBP	0,687	0,331	0,749	0,826	0,308	0,347	0,308	**0,827**	0,824	0,579	0,246
LCH	0,654	0,310	0,830	0,829	0,351	0,358	0,459	0,829	***0,831**	0,606	0,234
MOMENTS	0,351	0,308	0,451	0,730	0,305	0,306	0,306	**0,731**	0,507	0,444	0,178
MPO	0,474	0,341	0,478	0,792	0,303	0,304	0,305	0,805	**0,826**	0,514	0,230
MPOC	0,488	0,298	0,499	0,532	0,344	0,357	0,401	**0,535**	0,502	0,439	0,090
PHOG	0,739	0,353	0,830	0,825	0,349	0,355	0,420	0,826	***0,831**	0,614	0,235
RCS	0,459	0,317	**0,478**	0,387	0,322	0,376	0,360	0,388	**0,478**	0,396	0,062
TAMURA	0,801	0,350	0,829	0,829	0,277	0,274	0,414	0,829	**0,830**	0,604	0,264
MÉDIA	0,539	0,331	0,602	0,737	0,321	0,347	0,364	0,740	0,704	-	

Tabela 14 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor, considerando o algoritmo de agrupamento CURE e o conjunto D_6. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
ACC	0,71	0,70	0,76	459,62
BIC	0,74	0,74	0,61	272,83
CEDD	0,74	0,71	0,67	266,75
FCTH	0,73	0,70	0,59	661,17
GABOR	0,47	0,47	**0,42**	2555,35
GCH	**0,77**	**0,78**	0,65	214,99
HARACOLOR	0,59	0,59	0,46	**12682,76**
HARAFULL	0,65	0,63	0,74	2668,87
HARALICK	0,53	0,55	0,43	3995,41
JCD	0,73	0,70	0,64	399,04
LBP	0,34	0,37	0,64	475,86
LCH	0,39	0,55	0,79	2,55
MOMENTS	0,72	0,73	0,65	2713,11
MPO	0,60	0,60	0,46	8487,48
MPOC	0,72	0,69	0,52	325,35
PHOG	0,40	0,55	0,71	5,98
RCS	0,72	0,73	0,59	2305,36
TAMURA	0,39	0,55	0,53	6,76

Tabela 15 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor, considerando o algoritmo de agrupamento CURE e o conjunto D_7. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
ACC	0,85	0,85	0,81	227,48
BIC	0,92	0,92	0,56	288,00
CEDD	0,92	0,92	0,93	132,14
FCTH	**0,93**	**0,92**	0,80	239,81
GABOR	0,70	0,74	**0,42**	2339,05
GCH	0,92	0,91	0,50	182,93
HARACOLOR	0,63	0,71	0,44	5414,58
HARAFULL	0,68	0,75	0,63	2047,40
HARALICK	0,63	0,64	0,44	2266,76
JCD	0,92	0,91	0,98	182,74
LBP	0,41	0,43	0,56	486,93
LCH	0,51	0,66	0,61	3,14
MOMENTS	0,68	0,75	0,56	2532,23
MPO	0,58	0,67	0,50	**5585,58**
MPOC	0,80	0,72	0,54	171,34
PHOG	0,51	0,65	0,78	5,27
RCS	0,92	0,92	0,63	868,74
TAMURA	0,51	0,66	0,58	4,79

Tabela 16 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor, considerando o algoritmo de agrupamento ROCK e o conjunto D_8. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
ACC	0,78	0,75	0,93	185,74
BIC	0,43	0,59	0,47	4,22
CEDD	0,44	0,59	1,05	1,17
FCTH	0,43	0,59	0,61	2,77
GABOR	0,43	0,42	0,39	2125,31
GCH	0,44	0,59	0,52	3,71
HARACOLOR	0,46	0,51	0,40	**8150,98**
HARAFULL	0,46	0,54	0,68	4103,04
HARALICK	0,43	0,58	0,26	19,61
JCD	0,43	0,59	0,67	2,43
LBP	0,43	0,59	0,76	22,62
LCH	0,44	0,59	1,05	0,93
MOMENTS	0,46	0,52	0,65	4752,47
MPO	0,45	0,49	0,41	7668,71
MPOC	0,45	0,47	0,57	1379,66
PHOG	0,44	0,59	0,45	4,65
RCS	**0,81**	**0,81**	0,45	1268,49
TAMURA	0,44	0,59	**0,37**	7,89

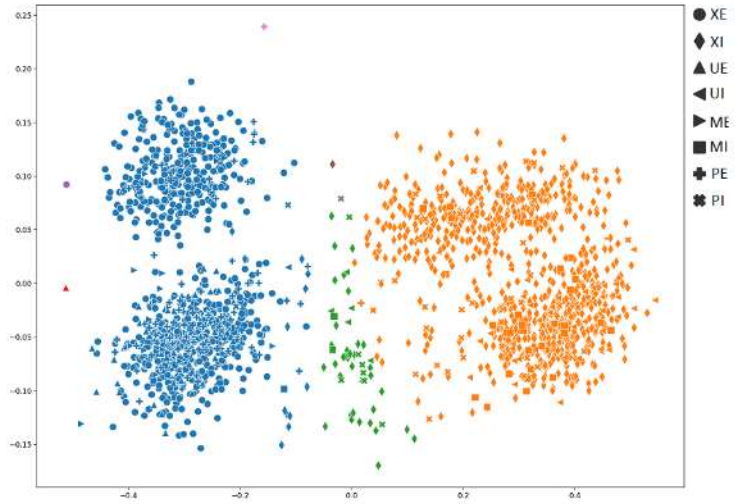
Tabela 17 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor, considerando o algoritmo de agrupamento ROCK e o conjunto D_9. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
ACC	0,78	0,75	1,23	238,77
BIC	0,44	0,59	0,47	4,34
CEDD	0,44	0,59	1,02	1,30
FCTH	0,43	0,59	0,61	2,76
GABOR	0,44	0,43	0,40	2624,93
GCH	0,44	0,59	0,51	3,93
HARACOLOR	0,47	0,56	0,37	5585,53
HARAFULL	0,47	0,55	0,65	4675,58
HARALICK	0,43	0,58	0,26	19,61
JCD	0,43	0,59	0,63	2,72
LBP	0,43	0,59	0,81	29,00
LCH	0,44	0,59	0,85	1,44
MOMENTS	0,47	0,56	0,60	4486,74
MPO	0,48	0,53	0,38	**6614,24**
MPOC	0,46	0,48	0,58	1607,25
PHOG	0,44	0,59	0,44	4,85
RCS	**0,81**	**0,81**	0,46	1629,90
TAMURA	0,44	0,59	**0,35**	8,64

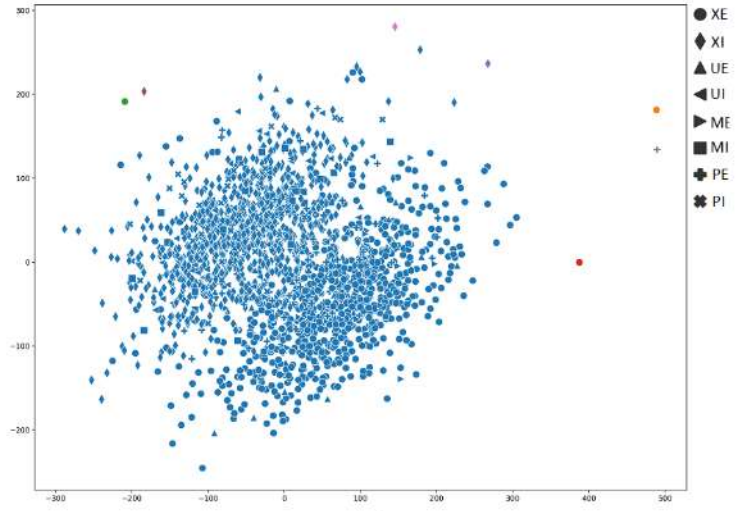
Tabela 18 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor, considerando o algoritmo de agrupamento ROCK e o conjunto D_{10}. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
ACC	0,46	0,59	0,93	556,89
BIC	0,83	0,84	0,45	4,78
CEDD	0,83	0,83	0,74	1,76
FCTH	0,83	0,83	0,64	2,78
GABOR	0,67	0,68	0,34	1251,93
GCH	0,83	0,84	0,46	5,01
HARACOLOR	0,64	0,65	0,34	1107,21
HARAFULL	0,30	0,42	0,47	**7080,66**
HARALICK	0,82	0,82	0,26	19,61
JCD	0,83	0,83	0,55	3,19
LBP	0,82	0,83	0,37	67,47
LCH	**0,83**	**0,84**	0,94	1,18
MOMENTS	0,51	0,54	0,54	2663,56
MPO	0,83	0,83	0,41	15,28
MPOC	0,50	0,59	0,48	1332,25
PHOG	0,83	0,84	0,42	5,23
RCS	0,48	0,59	0,56	3197,07
TAMURA	0,83	0,84	**0,29**	12,21

Figura 5 – Representação gráfica dos agrupamentos obtidos com a técnica PCA para o conjunto D_9, considerando a melhor combinação (par descritor e algoritmo de agrupamento) de acordo com cada métrica (acurácia, DAVIES e CALINSKI, respectivamente): (a) RCS-ROCK. (b) TAMURA-ROCK. (c) MPO-ROCK.



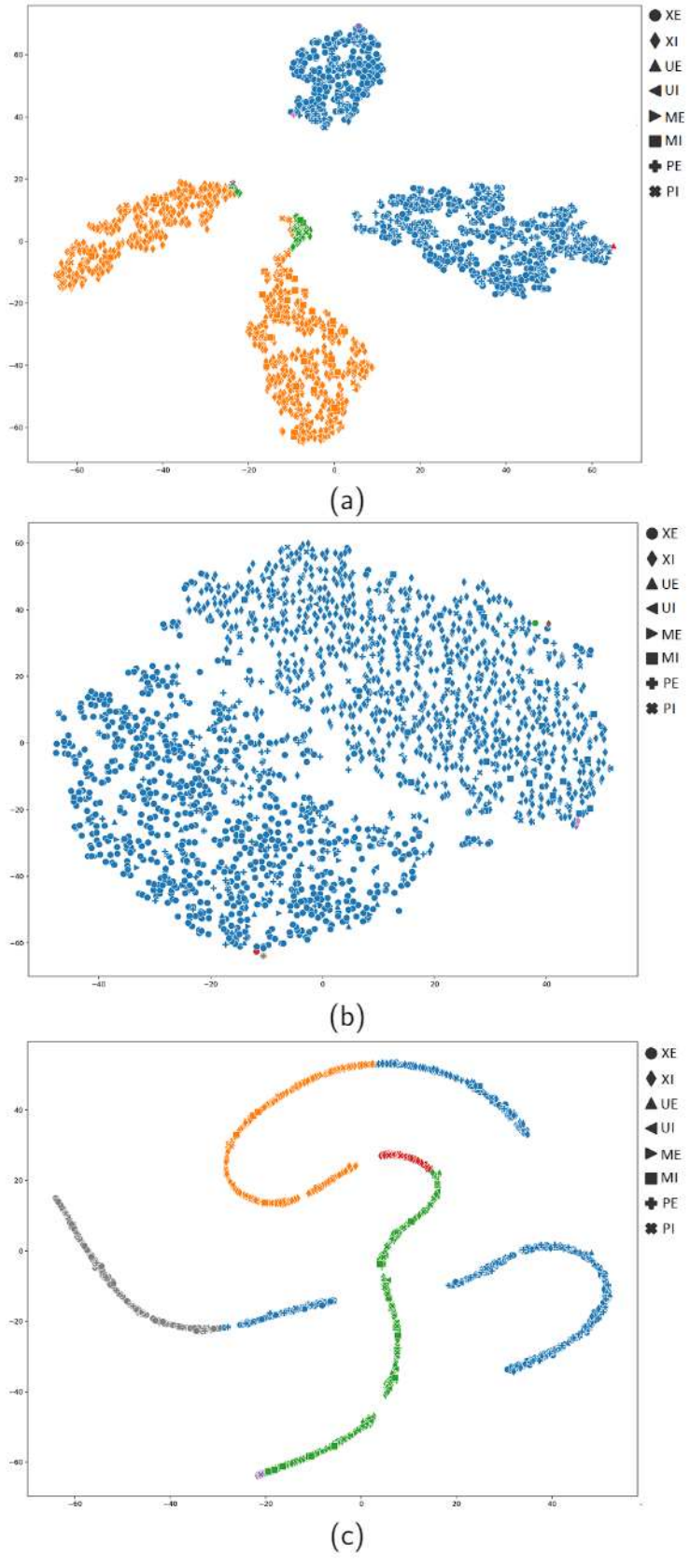
(a)



(b)

(c)

Figura 6 – Representação gráfica dos agrupamentos obtidos com a técnica TSNE para o conjunto D_9, considerando a melhor combinação (par descritor e algoritmo de agrupamento) de acordo com cada métrica (acurácia, DAVIES e CALINSKI, respectivamente): (a) RCS-ROCK. (b) TAMURA-ROCK. (c) MPO-ROCK.



4 CONCLUSÃO

O presente trabalho teve como objetivo realizar uma avaliação experimental extensiva, considerando diferentes técnicas de aprendizado não supervisionadas aplicadas a conjuntos de imagens de sementes de soja submetidas ao teste de tetrazólio. Para tanto, foram utilizados (5) conjuntos de imagens, considerando diferentes cenários, incluindo diferentes danos e/ou seus respectivos níveis. Para a descrição de tais imagens, foram considerados (18) diferentes extratores de características tradicionais. Foram avaliados 9 algoritmos de agrupamento, entre eles particionais, hierárquicos e baseados em densidade. Para validação foram consideradas 4 métricas (acurácia, FOWLKES, DAVIES e CALINSKI) e 2 técnicas de redução de dimensionalidade (PCA e TSNE).

A partir dos resultados obtidos, o mesmo comportamento foi observado utilizando os diferentes conjuntos de dados, em que as melhores acurácias foram obtidas de acordo com a quantidade de amostras perfeitas (sem dano) existentes no conjunto. O desbalanceamento de amostras em cada classe, incluindo o excesso de amostras perfeitas atrapalham o desempenho dos algoritmos de agrupamento, especialmente referente aos agrupamentos de classes que apresentam poucas amostras.

De maneira geral os melhores resultados de acurácia por cada descritor foram atingidos com os algoritmos de agrupamento AGNES, CURE e ROCK. Sendo que, dentre todos os descritores analisados, FCTH, GCH e RCS atingiram os maiores valores de acurácia, com as combinações FCTH-CURE, GCH-CURE e RCS-ROCK.

Com relação às outras métricas, a partir da análise dos gráficos gerados, FOWLKES apresentou resultados similares aos da acurácia, isso sugere que essa métrica pode ser utilizada como substituta da acurácia para análise de agrupamentos. Os melhores resultados para a métrica DAVIES foram agrupamentos em que quase todas as amostras pertenciam a um grupo, indicando que essa métrica não deve ser utilizada para representar bons agrupamentos. Para a métrica CALINSKI observou-se em seus melhores resultados, as amostras de classes perfeitas, que correspondem a maioria das amostras, foram divididas em múltiplos grupos, logo a acurácia em todos esses casos foram baixas, indicando que essa métrica também não deve ser utilizada para representar bons agrupamentos.

Referentes às técnicas de redução de dimensionalidade analisadas, PCA e TSNE não conseguiram agrupar adequadamente todas as classes de danos e/ou níveis, exceto a classe perfeita. Em uma comparação entre as duas técnicas os resultados foram mais coesos e separados com a técnica TSNE, entretanto quando utilizada com determinados descritores foi apresentada uma grande sobreposição entre as amostras (Figura 6 (c)).

De maneira geral, os agrupamentos não conseguiram identificar de maneira eficaz as diferentes classes de danos, somente as porções perfeitas. Isso pode ter ocorrido, porque em todos os conjuntos mais de 80% das amostras são da classe perfeita. Melhorias nos agrupamentos

podem ser obtidas utilizando outros descritores que melhor caracterizem as classes de danos e níveis. Apesar dos resultados obtidos, o presente trabalho apresenta contribuições significativas, dado que possibilita identificar os descritores e algoritmos de agrupamento a serem utilizados como pré-processamento em outras abordagens de aprendizado. Sendo útil, por exemplo, em muitas estratégias de aprendizado ativo, as quais requerem uma pré-organização dos dados eficaz, de forma a acelerar o processo de aprendizado e de classificação.

Referências

ALVES, D. H. A. **Técnicas de Aprendizado Profundo e Ativo Para Classificação de Bioimagens**. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2018. Citado 2 vezes nas páginas 1 e 8.

ANKERSTM, M.; BREUNING, M. M. Optics: Ordering points to identify the clustering structure. **ACM SIGMOD international conference on management of data**, p. 49–60, 1999. Citado na página 5.

BEZDEK, J. C.; PAL, S. K. Fuzzy models for pattern recognition : methods tha search for structures in data. **IEEE Press**, 1992. Citado na página 5.

BOSCH, A.; ZISSERMAN, A.; MUNOZ, X. Representing shape with a spatial pyramid kernel. **Proceedings of the 6th ACM international conference on image and video retrieval**, p. 401–408, 2007. Citado na página 3.

BRESSAN, R. S. **Aprendizado Ativo para Recuperação e Classificação de Imagens**. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2018. Citado 3 vezes nas páginas 1, 8 e 13.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis,communications. **Statistics - Theory and Methods**, v. 1, n. 3, p. 1–27, 1974. Citado 2 vezes nas páginas 6 e 7.

CAMARGO, G. **Aprendizado Ativo para Classificação de Bioimagens**. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2017. Citado 2 vezes nas páginas 1 e 8.

CAMPBELL, J. B. Introduction to remote sensing. **Taylor and Francis**, 2002. Citado na página 4.

CHATZICHRISTOFIS, S. A.; BOUTALIS, Y. S. Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. **International Conference on Computer Vision Systems**, p. 312–322, 2008a. Citado na página 3.

CHATZICHRISTOFIS, S. A.; BOUTALIS, Y. S. Fcth: Fuzzy color and texture histogram-low level feature for accurate image retrieval. **Image Analysis for Multimedia Interactive Services**, p. 191–196, 2008b. Citado na página 3.

COSTA, A. F.; TRAINA, A. J. M. **Mineração de imagens médicas utilizando características de forma**. 2012. Citado na página 3.

DAVIES, D. L.; BOLDIN, D. W. A cluster separation measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, PAMI-1, n. 2, p. 224–227, 1979. Citado 2 vezes nas páginas 6 e 7.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. **KKD**, v. 96, p. 226–231, 1996. Citado 2 vezes nas páginas 5 e 14.

FOWLKES, E. B.; MALLOWS, C. L. A method for comparing two hierarchical clusterings. **Journal of the American Statistical Association**, 1983. Citado na página 6.

FREITAS, C. M. A cultura da soja no brasil: O crescimento da produção brasileira e o surgimento de uma nova fronteira agricola. **Enciclopa Biosfera**, 2011. Citado na página 1.

GISLASON, P. O.; BENEDIKTSSON, J.; SVEINSSON, J. R. Random forest for land cover classification. **Pattern Recognition Letters**, p. 294–300, 2006. Citado na página 4.

GONZALES, R. C.; WOODS, R. E. **Digital Image Processing**. [S.l.]: Pearson, 2017. Citado na página 3.

GRAF, F. 2015. Jfeaturelib v1.6.3. Citado na página 3.

GUHA, S.; RASTOGI, R.; SHIM, K. Cure: an efficient clustering algorithm for large databases. **ACM Sigmod International Conference on Management of Data**, p. 73–84, 1998. Citado na página 5.

GUHA, S.; RASTOGI, R.; SHIM, K. Rock: A robust clustering algorithm for categorical attributes. **Infomation Systems**, v. 25, p. 345–366, 2000. Citado na página 5.

GUO, Z.; ZHANG, L.; ZHANG, D. Rotation invariant texture classification using lbp variance (lbpv) with global matching. **PR**, p. 706–709, 2010. Citado na página 3.

HARALICK, R. M.; SHANMUGAM, K.; DISTEIN, I. Textural features for image classification. **IEEE Transactions on Systems, Man, and Cybernetics**, p. 610–621, 1973. Citado na página 3.

HUANG, J.; KUMAR, S. R.; ZABIH, R. Image indexing using color correlagrams. **IEEE Computer Society Conference**, p. 762–768, 1997. Citado na página 3.

KAUFMAN, L.; ROUSSEEUW, P. Agglometarive nesting. **New York: Wiley Inter-Science**, v. 1, p. 199–252, 1990a. Citado na página 5.

KAUFMAN, L.; RUSSEEUW, P. J. Clustering by means of medoids, in statical data analysis. **North-Holland**, 1987. Citado na página 5.

KIST, B. B. Anuário brasileiro da soja 2018. **Santa Cruz do Sul : Editora Gazeta Santa Cruz**, 2018. Citado na página 1.

MAATEN, L. J. P.; HINTON, G. E. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, p. 2579–2605, 2008. Citado na página 8.

MACQUEEN, J. B. Some methods for classification and analisys of multivariate observations. **Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Propabilty**, v. 1, 1967. Citado na página 5.

MORBIEU, S. **Accuracy: from classification to clustering evaluation**. 2019. <https://smorbieu.gitlab.io/accuracy-from-classification-to-clustering-evaluation/>. Acesso: 3 de agosto de 2020. Citado 2 vezes nas páginas 6 e 15.

MáXIMO O. A; FERNANDES, D. Classificação supervisionada de imagens sar do sivam pré-definidas. **XII Simpósio Brasileiro de Sensoriamento Remoto**, 2005. Citado na página 4.

NETO, F. et al. Metodologia teste de tetrazólio em semente de soja. **Embrapa Soja**, 2008. Citado na página 1.

NOVIKOV, A. PyClustering: Data mining library. **Journal of Open Source Software**, The Open Journal, v. 4, n. 36, p. 1230, apr 2019. Disponível em: <https://doi.org/10.21105/joss.01230>. Citado na página 14.

P., P. J.; FALCãO, A. X.; SUZUKI, C. T. N. Supervised pattern classification based on optimum-path. **Proceeding of the 4th International Symposium on Advances in Visual Computing**, p. 120–131, 2009. Citado na página 4.

PALIWAL, K. K. Dimensionality reduction of the enhanced feature set for the hmm-based speech recognizer. **Digital Signal Processing**, v. 2, p. 157–173, 1992. Citado na página 8.

PEREIRA, D. F. **Técnicas de Aprendizado Ativo para Avaliação do Vigor de Sementes de Soja**. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2019. Citado 4 vezes nas páginas 1, 8, 10 e 11.

PERPINAN, C. M. A. **Continuous Latent Variable Models for Dimensionality Reduction and Sequential Data Reconstruction**. Tese (Doutorado) — University of Sheffield, UK, 2 2001. Citado na página 8.

RAYMOND, N.; HAN, J. A method for clustering objects for spatial data mining. **IEE Trans. Knowledge and Data Engineering**, 2002. Citado na página 5.

RENDóN, E. Internal versus external cluster validation indexes. **International Journal Of Computers And Communications**, v. 5, n. 1, p. 27–33, 2011. Citado na página 7.

SANTANNA, M. G. F. **Recuperação e Classificação Automática do Vigor de Sementes de Soja**. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2014. Citado na página 1.

SMITH, J. R.; CHANG, S. F. Local color and texture extraction and spatial query. **Intl. Conference on Image Processing**, v. 3, p. 1011–1014, 1996. Citado na página 3.

SMITH, J. R.; CHANG, S. F. Evaluantion of multiple clustering solutions. **MultiClust@ ECML/PKDD**, p. 55–66, 2011. Citado na página 3.

STEHLING, R. O.; NASCIMENTO, M. A.; FALCÃO, A. X. A compact and efficient image retrieval approach based on border/interior pixel classification. **Intl. Conference on Information and Knowledge Management**, p. 102–109, 2002. Citado na página 3.

STRICKER, M. A.; OREGON, M. Similarity of color images. **Storage and Retrieval for Image and Video Databases**, v. 2420, n. III, p. 381–393, 1995. Citado na página 3.

SWAIN, M. J.; BALLARD, D. H. Color indexig. **International journal of computer vision**, 1991. Citado na página 3.

TAMURA, H.; MORI, S.; YAMAWAKI, T. Textural features corresponding to visual perception. **IEEE Transactions on Systems, man and cybernetics**, p. 460–473, 1978. Citado na página 3.

TIPPING, M. E.; BISHOP, C. M. Probabilistic principal component analysis. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 3, n. 61, p. 611–622, 1999. Citado na página 8.

VAPNIK, V.; GOLOWICH, S. E.; SMOLA, A. Support vector method for function approximation, regression estimation, and signal processing. **Neural Information Processing Systems**, v. 9, 1997. Citado na página 4.

ZHANG, D. et al. Content-based image retrieval using gabor texture features. **IEEE Transactions PAMI**, p. 13–15, 2000. Citado na página 3.

Apêndices

APÊNDICE A – Resultados de todas as métricas obtidas para o conjunto D6

Tabela 19 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento AGNES e o conjunto D_6. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	10	0,58	0,58	1,29	650,41
BIC	10	0,70	0,69	1,16	316,80
CEDD	10	0,63	0,61	1,74	298,55
FCTH	10	**0,72**	**0,72**	0,92	883,18
Gabor	10	0,40	0,42	0,44	3606,19
GCH	10	0,70	0,70	1,42	284,29
Haralick	10	0,47	0,50	0,53	**16692,26**
HaralickColor	10	0,58	0,58	0,77	4293,54
HaralickFull	10	0,54	0,56	**0,43**	3207,55
JCD	10	0,72	0,72	1,55	474,61
LBP	10	0,28	0,29	0,96	895,84
LCH	10	0,57	0,61	1,33	161,23
Moments	10	0,63	0,63	0,73	3785,94
MPO	10	0,61	0,61	0,46	7977,59
MPOC	10	0,60	0,58	0,65	465,75
PHOG	10	0,44	0,50	1,15	20,47
ReferenceColorSimilarity	10	0,59	0,61	0,90	2806,29
Tamura	10	0,54	0,48	1,92	83,59

Tabela 20 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento CLARANS e o conjunto D_6. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	10	0,29	0,37	1,72	830,16
BIC	10	0,37	0,42	1,36	631,49
CEDD	10	0,34	0,38	2,68	432,96
FCTH	10	0,42	**0,46**	1,81	1285,12
Gabor	10	0,28	0,33	**0,54**	5928,08
GCH	10	0,41	0,44	2,07	392,61
Haralick	10	0,29	0,37	0,57	18139,47
HaralickColor	10	0,38	0,42	1,02	5633,43
HaralickFull	10	0,26	0,34	0,56	6574,37
JCD	10	0,32	0,38	2,49	641,17
LBP	10	0,19	0,21	1,16	1410,23
LCH	10	0,37	0,42	2,49	240,16
Moments	10	**0,43**	0,45	0,99	6068,04
MPO	10	0,36	0,41	0,55	**19708,21**
MPOC	10	0,29	0,35	1,16	1336,99
PHOG	10	0,24	0,28	2,21	157,23
ReferenceColorSimilarity	10	0,37	0,41	1,36	4379,59
Tamura	10	0,28	0,35	2,12	176,78

Tabela 21 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento CURE e o conjunto D_6. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
ACC	10	0,71	0,70	0,76	459,62
BIC	10	0,74	0,74	0,61	272,83
CEDD	10	0,74	0,71	0,67	266,75
FCTH	10	0,73	0,70	0,59	661,17
GABOR	10	0,47	0,47	**0,42**	2555,35
GCH	10	**0,77**	**0,78**	0,65	214,99
HARACOLOR	10	0,59	0,59	0,46	**12682,76**
HARAFULL	10	0,65	0,63	0,74	2668,87
HARALICK	10	0,53	0,55	0,43	3995,41
JCD	10	0,73	0,70	0,64	399,04
LBP	10	0,34	0,37	0,64	475,86
LCH	10	0,39	0,55	0,79	2,55
MOMENTS	10	0,72	0,73	0,65	2713,11
MPO	10	0,60	0,60	0,46	8487,48
MPOC	10	0,72	0,69	0,52	325,35
PHOG	10	0,40	0,55	0,71	5,98
RCS	10	0,72	0,73	0,59	2305,36
TAMURA	10	0,39	0,55	0,53	6,76

Tabela 22 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento DBSCAN e o conjunto D_6. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
Auto Color C.	10	**0,70**	**0,66**	1,14	**234,83**
BIC	6	0,72	0,69	1,02	331,88
CEDD	5	0,39	0,55	1,07	3,98
FCTH	8	0,54	0,49	1,19	73,69
GABOR	5	0,39	0,55	**0,41**	27,34
GCH	7	**0,71**	**0,67**	0,87	229,22
HARALICK	10	0,39	0,47	**0,98**	**29,29**
HARALICK C	8	0,70	0,65	1,52	592,45
HARALICK F	7	0,39	0,51	1,65	13,28
JCD	6	0,39	0,55	1,31	2,77
LBP	7	0,39	0,54	0,84	9,14
LCH	5	0,58	0,62	1,24	282,05
MOMENTS	6	0,72	0,69	0,54	1500,92
MPO	9	0,55	0,50	8,16	378,28
MPOC	9	0,39	0,55	0,86	5,16
PHOG	7	0,39	0,55	0,80	6,78
REF. COL. S.	8	0,70	0,65	0,88	565,56
TAMURA	2	0,39	0,55	0,99	2,98

Tabela 23 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento FCM e o conjunto D_6. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	10	**0,31**	**0,37**	1,38	**1055,85**
BIC	10	0,27	0,36	1,66	657,37
CEDD	8	0,42	0,48	2,53	463,98
FCTH	9	0,35	0,44	2,06	1487,90
Gabor	10	0,26	0,32	**0,54**	7547,16
GCH	10	**0,32**	**0,40**	3,50	394,55
Haralick	10	0,35	0,42	**0,54**	**27036,81**
HaralickColor	10	0,32	0,40	0,98	7005,70
HaralickFull	10	0,30	0,37	0,52	9715,45
JCD	10	**0,40**	**0,46**	3,41	535,86
LBP	10	0,17	0,19	1,21	1538,74
LCH	8	0,41	0,48	3,00	271,34
Moments	10	0,27	0,37	0,90	7509,44
MPO	10	0,32	0,39	**0,52**	25774,45
MPOC	10	0,26	0,34	0,92	1664,17
PHOG	10	0,36	0,35	4,41	121,64
ReferenceColorSimilarity	10	0,27	0,37	1,19	5173,29
Tamura	10	0,28	0,34	3,96	169,63

Tabela 24 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento KMEANS e o conjunto D_6. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	10	0,34	0,41	1,23	1077,50
BIC	10	0,32	0,40	1,33	710,35
CEDD	10	0,33	0,43	2,21	472,43
FCTH	10	**0,45**	**0,53**	1,47	1506,71
Gabor	10	0,28	0,34	0,51	7023,69
GCH	10	0,32	0,40	1,70	479,47
Haralick	10	0,37	0,43	0,53	**26609,13**
HaralickColor	10	0,36	0,43	0,96	6823,77
HaralickFull	10	0,32	0,39	**0,50**	9362,58
JCD	10	**0,39**	**0,47**	2,07	700,18
LBP	10	0,19	0,21	1,07	1562,91
LCH	10	0,34	0,43	2,19	262,44
Moments	10	0,32	0,40	0,93	7297,98
MPO	10	0,33	0,40	**0,52**	25020,60
MPOC	10	0,30	0,36	0,94	1669,17
PHOG	10	0,28	0,31	1,79	184,79
ReferenceColorSimilarity	10	0,35	0,42	1,18	4765,36
Tamura	10	0,28	0,35	1,71	219,51

Tabela 25 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento OPTICS e o conjunto D_6. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	10	0,49	0,59	1,42	460,79
BIC	10	0,50	0,56	1,49	417,41
CEDD	7	0,66	0,64	2,00	415,63
FCTH	5	**0,64**	**0,65**	1,50	1622,03
Gabor	10	0,29	0,34	**0,52**	7090,84
GCH	10	**0,54**	**0,60**	1,88	297,33
Haralick	10	0,36	0,42	0,52	**20857,38**
HaralickColor	10	0,38	0,45	0,91	4925,45
HaralickFull	10	0,30	0,36	**0,54**	8527,32
JCD	5	**0,65**	**0,66**	1,81	951,47
LBP	10	0,20	0,21	1,09	1296,48
LCH	6	0,60	0,66	1,58	310,33
Moments	10	0,36	0,45	0,86	5653,14
MPO	10	0,33	0,39	**0,52**	**21950,32**
MPOC	10	0,30	0,37	1,00	1480,28
PHOG	6	0,53	0,48	1,95	163,72
ReferenceColorSimilarity	10	0,48	0,57	0,96	3171,62
Tamura	8	0,50	0,52	1,79	146,15

Tabela 26 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento OPTICS e o conjunto D_6. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
Auto Color C.	10	**0,70**	**0,66**	1,14	**231,38**
BIC	6	0,72	0,69	1,00	327,61
CEDD	5	0,39	0,55	1,07	3,98
FCTH	8	**0,54**	**0,49**	1,13	68,61
GABOR	5	0,39	0,55	**0,41**	27,34
GCH	7	**0,71**	**0,67**	0,82	222,22
HARALICK	10	0,39	0,48	**0,98**	**26,93**
HARALICK C	8	0,70	0,65	1,55	577,54
HARALICK F	7	0,40	0,52	**1,66**	7,75
JCD	6	**0,39**	**0,55**	1,31	2,77
LBP	7	0,39	0,54	0,84	9,14
LCH	5	0,58	0,62	1,16	278,71
MOMENTS	6	0,72	0,69	0,52	1445,70
MPO	9	0,55	0,50	**9,31**	**375,40**
MPOC	9	0,39	0,55	0,86	5,16
PHOG	7	0,39	0,55	0,80	6,78
REF. COL. S.	8	0,70	0,65	0,88	558,89
TAMURA	2	0,39	0,55	0,99	2,98

Tabela 27 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento ROCK e o conjunto D_6. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	10	**0,75**	**0,73**	0,47	**351,22**
BIC	10	0,39	0,55	0,47	4,62
CEDD	10	0,39	0,56	0,87	1,51
FCTH	10	**0,72**	**0,68**	0,47	617,54
Gabor	10	0,55	0,51	**0,37**	1223,28
GCH	10	**0,39**	**0,55**	0,58	2,86
Haralick	10	0,73	0,73	**0,39**	**3358,01**
HaralickColor	10	0,72	0,72	0,65	2561,76
HaralickFull	11	0,39	0,54	**0,31**	22,99
JCD	10	**0,39**	**0,56**	0,86	1,49
LBP	10	0,40	0,53	0,95	48,88
LCH	10	0,39	0,56	1,15	0,80
Moments	10	0,72	0,73	0,63	2743,04
MPO	10	0,68	0,68	**0,41**	**4630,57**
MPOC	10	0,41	0,53	0,77	27,93
PHOG	10	0,40	0,56	0,51	4,22
ReferenceColorSimilarity	10	**0,75**	**0,74**	0,48	2034,74
Tamura	10	0,39	0,55	0,39	6,87

APÊNDICE B – Resultados de todas as métricas obtidas para o conjunto D7

Tabela 28 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento AGNES e o conjunto D_7. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	7	0,78	0,76	1,66	170,03
BIC	7	0,68	0,76	0,88	490,95
CEDD	7	0,75	0,78	1,24	189,12
FCTH	7	0,65	0,65	1,13	501,08
Gabor	7	0,66	0,69	0,47	2943,43
GCH	7	**0,78**	0,78	0,96	239,26
Haralick	7	0,57	0,67	0,54	**6122,97**
HaralickColor	7	0,59	0,67	0,70	1868,50
HaralickFull	7	0,63	0,64	0,50	2249,24
JCD	7	0,75	0,76	1,33	270,44
LBP	7	0,42	0,38	0,62	678,38
LCH	7	0,74	**0,82**	1,03	150,04
Moments	7	0,61	0,69	0,84	2554,03
MPO	7	0,64	0,71	**0,46**	3654,31
MPOC	7	0,54	0,53	0,88	239,00
PHOG	7	0,60	0,51	1,56	47,72
ReferenceColorSimilarity	7	0,58	0,66	0,81	1820,62
Tamura	7	0,56	0,52	1,63	96,94

Tabela 29 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento CLARANS e o conjunto D_7. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	7	0,52	0,60	1,54	358,87
BIC	7	0,47	0,57	1,48	693,15
CEDD	7	0,52	0,53	2,56	246,83
FCTH	7	**0,59**	**0,66**	1,68	791,93
Gabor	7	0,40	0,52	0,57	5352,88
GCH	7	0,47	0,55	1,85	311,29
Haralick	7	0,43	0,55	0,56	**7754,26**
HaralickColor	7	0,37	0,55	0,93	3605,70
HaralickFull	7	0,46	0,51	**0,52**	3059,07
JCD	7	0,41	0,59	1,90	435,52
LBP	7	0,24	0,28	1,06	994,77
LCH	7	0,47	0,50	2,12	206,62
Moments	7	0,35	0,59	0,92	3994,58
MPO	7	0,42	0,53	0,58	7249,56
MPOC	7	0,41	0,47	1,16	586,29
PHOG	7	0,40	0,42	1,82	138,31
ReferenceColorSimilarity	7	0,41	0,56	0,93	2788,73
Tamura	7	0,37	0,46	2,25	139,01

Tabela 30 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento CURE e o conjunto D_7. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
ACC	7	0,85	0,85	0,81	227,48
BIC	7	0,92	0,92	0,56	288,00
CEDD	7	0,92	0,92	0,93	132,14
FCTH	7	**0,93**	**0,92**	0,80	239,81
GABOR	7	0,70	0,74	**0,42**	2339,05
GCH	7	0,92	0,91	0,50	182,93
HARACOLOR	7	0,63	0,71	0,44	5414,58
HARAFULL	7	0,68	0,75	0,63	2047,40
HARALICK	7	0,63	0,64	0,44	2266,76
JCD	7	0,92	0,91	0,98	182,74
LBP	7	0,41	0,43	0,56	486,93
LCH	7	0,51	0,66	0,61	3,14
MOMENTS	7	0,68	0,75	0,56	2532,23
MPO	7	0,58	0,67	0,50	**5585,58**
MPOC	7	0,80	0,72	0,54	171,34
PHOG	7	0,51	0,65	0,78	5,27
RCS	7	0,92	0,92	0,63	868,74
TAMURA	7	0,51	0,66	0,58	4,79

Tabela 31 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento DBSCAN e o conjunto D_7. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	5	0,84	0,78	0,98	158,63
BIC	6	0,88	0,85	0,71	307,42
CEDD	2	0,79	0,70	1,40	300,38
FCTH	4	0,51	0,66	1,06	3,99
Gabor	7	**0,85**	**0,85**	**0,54**	**1038,79**
GCH	3	0,87	0,83	1,04	315,76
Haralick	7	0,47	0,56	0,99	22,36
HaralickColor	3	0,84	0,78	0,54	599,12
HaralickFull	7	0,64	0,57	0,60	102,18
JCD	2	0,87	0,81	0,97	608,50
LBP	3	0,50	0,65	0,70	12,52
LCH	5	0,53	0,56	1,89	15,38
Moments	5	0,84	0,79	1,04	442,23
MPO	7	0,55	0,54	1,26	21,42
MPOC	3	0,52	0,65	0,57	19,92
PHOG	2	0,52	0,66	1,20	4,72
ReferenceColorSimilarity	3	0,84	0,78	0,56	689,93
Tamura	4	0,51	0,65	0,71	6,37

Tabela 32 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento FCM e o conjunto D_7. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	7	0,40	0,51	1,30	441,92
BIC	7	0,41	0,53	1,33	271,70
CEDD	7	0,44	0,55	3,33	897,11
FCTH	7	**0,54**	**0,65**	1,58	5709,63
Gabor	7	0,42	0,52	0,55	353,66
GCH	7	0,40	0,52	1,97	**9051,15**
Haralick	7	0,42	0,54	0,59	3933,24
HaralickColor	7	0,37	0,52	0,87	4792,30
HaralickFull	7	0,39	0,51	**0,52**	448,73
JCD	7	0,40	0,54	2,44	1353,50
LBP	7	0,24	0,29	0,78	232,76
LCH	6	0,47	0,60	2,03	4709,91
Moments	7	0,39	0,52	0,91	8648,03
MPO	7	0,43	0,54	0,55	667,17
MPOC	7	0,34	0,45	0,95	124,23
PHOG	7	0,37	0,39	2,59	2945,97
ReferenceColorSimilarity	7	0,35	0,51	1,11	157,21
Tamura	9	0,40	0,43	3,43	122,93

Tabela 33 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento KMEANS e o conjunto D_7. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	7	0,52	0,59	1,14	421,30
BIC	7	0,46	0,56	1,23	765,29
CEDD	7	0,47	0,58	1,94	290,70
FCTH	7	**0,57**	**0,67**	1,26	930,43
Gabor	7	0,43	0,53	0,54	5491,84
GCH	7	0,39	0,53	1,54	369,54
Haralick	7	0,43	0,55	0,57	**8806,52**
HaralickColor	7	0,41	0,54	0,81	3948,45
HaralickFull	7	0,39	0,51	**0,50**	4802,80
JCD	7	0,52	0,63	1,68	470,35
LBP	7	0,26	0,30	0,75	1348,94
LCH	7	0,41	0,56	1,83	213,88
Moments	7	0,41	0,53	0,87	4694,11
MPO	7	0,43	0,54	0,54	8395,98
MPOC	7	0,38	0,47	0,92	649,24
PHOG	7	0,34	0,40	1,57	149,63
ReferenceColorSimilarity	7	0,45	0,56	1,04	2758,87
Tamura	7	0,40	0,43	1,59	146,16

Tabela 34 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento KMEDOIDS e o conjunto D_7. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	7	0,44	0,57	1,48	313,14
BIC	7	**0,61**	**0,68**	1,34	502,71
CEDD	4	0,78	0,80	1,88	294,16
FCTH	5	0,61	0,68	1,45	710,75
Gabor	7	0,48	0,58	0,57	4234,96
GCH	6	0,74	0,80	1,69	271,92
Haralick	7	0,43	0,53	0,56	**8093,23**
HaralickColor	7	0,41	0,54	0,76	2859,52
HaralickFull	7	0,36	0,47	**0,53**	3969,43
JCD	6	0,63	0,69	2,08	362,84
LBP	7	0,29	0,32	1,08	878,37
LCH	6	0,74	0,81	1,58	180,59
Moments	7	0,42	0,58	0,98	2977,95
MPO	7	0,38	0,51	0,57	7221,94
MPOC	7	0,44	0,51	0,95	579,91
PHOG	7	0,42	0,42	1,88	118,88
ReferenceColorSimilarity	7	0,41	0,56	0,97	2525,75
Tamura	7	0,44	0,53	1,84	152,79

Tabela 35 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento OPTICS e o conjunto D_7. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	5	0,84	0,78	0,88	146,47
BIC	6	0,88	0,85	0,57	288,63
CEDD	2	0,85	0,79	1,19	430,18
FCTH	4	0,51	0,66	1,06	3,99
Gabor	7	**0,85**	**0,85**	**0,54**	**1038,79**
GCH	3	0,87	0,83	1,04	315,76
Haralick	7	0,52	0,56	1,00	21,18
HaralickColor	3	0,86	0,81	0,55	670,54
HaralickFull	7	0,64	0,57	0,60	98,21
JCD	2	0,86	0,80	0,99	581,09
LBP	3	0,49	0,63	0,59	52,74
LCH	3	0,58	0,57	2,22	22,32
Moments	5	0,84	0,79	1,10	409,79
MPO	7	0,68	0,56	1,32	17,53
MPOC	3	0,52	0,65	0,68	16,23
PHOG	2	0,57	0,63	1,22	63,96
ReferenceColorSimilarity	4	0,86	0,82	1,31	612,79
Tamura	4	0,51	0,65	0,71	6,37

Tabela 36 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento ROCK e o conjunto D_7. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	7	0,50	0,65	0,62	2,55
BIC	7	0,51	0,66	0,54	3,48
CEDD	7	0,51	0,66	1,17	0,79
FCTH	7	0,50	0,65	0,56	3,02
GABOR	7	0,76	0,76	0,43	2135,64
GCH	7	0,50	0,65	0,58	3,22
HARALICK	7	0,90	0,90	**0,34**	1065,62
HARALICK C	7	0,89	0,88	0,58	980,88
HARALICK F	7	0,61	0,65	0,44	2251,28
JCD	7	0,51	0,66	0,90	2,22
LBP	7	0,51	0,65	0,65	24,92
LCH	7	0,01	0,66	0,61	3,14
MOMENTS	7	0,82	0,82	0,60	1970,31
MPO	7	0,64	0,70	0,39	**4055,22**
MPOC	7	0,54	0,62	0,64	26,04
PHOG	7	0,51	0,66	0,47	4,35
REF. COL. S.	7	**0,92**	**0,91**	0,47	878,75
TAMURA	7	0,51	0,66	0,43	6,05

APÊNDICE C – Resultados de todas as métricas obtidas para o conjunto D8

Tabela 37 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento AGNES e o conjunto D_8. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	10	0,55	0,56	1,22	595,77
BIC	10	0,50	0,53	0,91	653,02
CEDD	10	**0,74**	0,68	1,34	238,10
FCTH	10	0,68	0,68	1,16	898,26
Gabor	10	0,33	0,32	0,52	9195,13
GCH	10	0,62	0,63	0,96	331,04
Haralick	10	0,36	0,42	0,53	**12365,94**
HaralickColor	10	0,43	0,47	0,75	4855,59
HaralickFull	10	0,52	0,51	**0,45**	4298,08
JCD	10	0,68	**0,68**	1,36	526,71
LBP	10	0,26	0,28	0,84	1077,48
LCH	10	0,63	0,66	1,12	135,41
Moments	10	0,45	0,48	0,74	4583,92
MPO	10	0,41	0,45	0,48	11161,58
MPOC	10	0,39	0,43	0,74	1908,19
PHOG	10	0,48	0,50	1,41	44,08
ReferenceColorSimilarity	10	0,49	0,52	0,93	2705,22
Tamura	10	0,55	0,49	2,10	132,70

Tabela 38 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento CLARANS e o conjunto D_8. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	10	0,32	0,40	1,43	979,73
BIC	10	**0,41**	**0,46**	1,40	894,26
CEDD	10	0,32	0,38	2,49	556,45
FCTH	10	0,39	0,44	1,59	1369,09
Gabor	10	0,30	0,32	0,55	10356,86
GCH	10	0,32	0,40	2,02	650,74
Haralick	10	0,28	0,39	0,53	**23660,28**
HaralickColor	10	0,34	0,43	0,86	6854,53
HaralickFull	10	0,35	0,40	0,55	10285,36
JCD	10	0,30	0,40	2,39	808,71
LBP	10	0,18	0,22	1,04	1651,85
LCH	10	0,34	0,43	2,51	336,70
Moments	10	0,29	0,39	0,85	8216,87
MPO	10	0,27	0,37	**0,53**	21089,76
MPOC	10	0,29	0,36	1,04	3930,77
PHOG	10	0,30	0,35	2,02	254,60
ReferenceColorSimilarity	10	0,33	0,41	1,07	4477,60
Tamura	10	0,26	0,35	2,63	286,20

Tabela 39 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento CURE e o conjunto D_8 Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	10	0,65	0,69	0,75	419,43
BIC	10	0,49	0,56	0,67	561,30
CEDD	10	0,44	0,59	0,96	1,58
FCTH	10	0,78	0,75	0,67	562,61
Gabor	10	0,32	0,34	0,47	8198,96
GCH	10	0,64	0,57	0,66	112,15
Haralick	10	0,33	0,43	0,49	**20543,96**
HaralickColor	10	0,47	0,51	0,64	4432,19
HaralickFull	10	0,40	0,45	**0,44**	6980,06
JCD	10	0,43	0,59	1,03	2,59
LBP	10	0,38	0,37	0,69	587,61
LCH	10	0,44	0,59	0,56	3,53
Moments	10	0,47	0,52	0,60	4755,63
MPO	10	0,36	0,42	0,50	15934,56
MPOC	10	0,46	0,48	0,50	1274,20
PHOG	10	0,44	0,59	0,56	5,32
ReferenceColorSimilarity	10	**0,81**	**0,81**	0,56	1247,89
Tamura	10	0,44	0,59	0,64	9,76

Tabela 40 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento DBSCAN e o conjunto D_8. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
Auto Color C.	6	0,58	0,53	1,15	94,45
BIC	8	0,43	0,58	0,96	4,18
CEDD	6	0,43	0,59	1,12	3,37
FCTH	6	0,43	0,59	1,38	4,12
GABOR	9	0,49	0,52	0,44	267,12
GCH	10	**0,64**	**0,68**	0,81	283,33
HARALICK	10	0,46	0,55	1,07	19,34
HARALICK C	8	0,42	0,45	1,21	665,11
HARALICK F	9	0,52	0,54	0,74	136,48
JCD	7	0,43	0,59	1,00	3,91
LBP	6	0,44	0,59	0,60	15,15
LCH	8	0,43	0,59	1,11	2,93
MOMENTS	10	0,46	0,45	1,94	571,48
MPO	9	0,47	0,52	0,99	47,48
MPOC	10	0,42	0,41	**0,77**	412,21
PHOG	8	0,44	0,58	0,88	7,17
REF. COL. S.	10	0,47	0,53	1,01	**1156,34**
TAMURA	7	0,44	0,59	0,80	6,42

Tabela 41 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento FCM e o conjunto D_8. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	10	0,31	0,39	1,33	1140,45
BIC	10	0,28	0,38	1,32	1125,92
CEDD	9	0,36	0,44	2,37	642,99
FCTH	10	0,34	**0,44**	1,63	1702,12
Gabor	10	0,22	0,28	0,54	16000,98
GCH	10	0,31	0,42	2,91	601,55
Haralick	10	0,29	0,39	0,52	**26096,53**
HaralickColor	10	0,28	0,39	0,84	7666,99
HaralickFull	10	0,29	0,37	0,55	14034,22
JCD	8	0,34	0,43	1,82	986,77
LBP	10	0,16	0,21	0,88	1952,47
LCH	5	0,51	0,58	1,62	534,34
Moments	10	0,25	0,37	0,91	10025,50
MPO	10	0,27	0,37	**0,52**	24461,21
MPOC	10	0,26	0,36	1,00	4325,60
PHOG	10	**0,37**	0,37	4,53	180,59
ReferenceColorSimilarity	10	0,27	0,39	1,06	5217,34
Tamura	10	0,28	0,36	4,12	266,67

Tabela 42 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento KMEANS e o conjunto D_8. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	10	0,34	0,41	1,21	1128,70
BIC	10	0,31	0,41	1,26	1133,85
CEDD	10	0,33	0,43	1,92	655,95
FCTH	10	**0,40**	**0,49**	1,32	1732,25
Gabor	10	0,23	0,29	0,54	15490,17
GCH	10	0,34	0,42	1,56	716,98
Haralick	10	0,28	0,38	0,53	**25824,26**
HaralickColor	10	0,31	0,41	0,80	7589,07
HaralickFull	10	0,30	0,39	**0,51**	13507,85
JCD	10	0,36	0,46	1,69	922,93
LBP	10	0,17	0,21	0,86	1975,92
LCH	10	0,37	0,45	1,71	393,89
Moments	10	0,27	0,37	0,90	9933,52
MPO	10	0,27	0,37	0,52	23820,41
MPOC	10	0,28	0,37	0,94	4341,72
PHOG	10	0,26	0,32	1,75	283,38
ReferenceColorSimilarity	10	0,32	0,41	1,06	4819,10
Tamura	10	0,27	0,36	1,73	349,92

Tabela 43 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento KMEDOIDS e o conjunto D_8. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	10	**0,58**	**0,64**	1,47	459,63
BIC	10	0,41	0,47	1,35	921,92
CEDD	6	0,48	0,52	1,73	795,66
FCTH	6	0,48	0,52	1,74	1692,32
Gabor	10	0,23	0,29	0,54	15708,50
GCH	8	0,44	0,53	1,70	614,66
Haralick	10	0,28	0,37	0,55	21212,62
HaralickColor	10	0,35	0,44	0,79	5419,91
HaralickFull	10	0,30	0,37	**0,52**	13729,58
JCD	10	0,41	0,48	2,11	657,88
LBP	10	0,17	0,21	1,11	1585,24
LCH	6	0,62	0,67	2,01	278,88
Moments	10	0,27	0,38	0,86	9379,62
MPO	10	0,26	0,36	0,54	**21357,40**
MPOC	10	0,35	0,40	0,91	3699,71
PHOG	9	0,45	0,44	1,97	222,72
ReferenceColorSimilarity	10	0,45	0,51	1,02	3728,38
Tamura	7	0,54	0,52	2,19	254,91

Tabela 44 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento OPTICS e o conjunto D_8. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	6	0,58	0,53	1,15	93,89
BIC	8	0,43	0,58	0,96	4,18
CEDD	6	0,43	0,59	1,12	3,37
FCTH	6	0,43	0,59	1,28	3,45
Gabor	9	0,49	0,52	0,43	256,95
GCH	10	**0,64**	**0,68**	0,81	283,33
Haralick	10	0,46	0,55	1,09	16,87
HaralickColor	8	0,42	0,45	1,21	659,42
HaralickFull	9	0,52	0,54	0,75	134,40
JCD	7	0,43	0,59	1,00	3,91
LBP	6	0,44	0,59	0,60	15,15
LCH	8	0,43	0,59	1,11	2,93
Moments	10	0,46	0,45	1,91	555,29
MPO	9	0,47	0,53	1,00	44,49
MPOC	10	0,42	0,41	**0,76**	408,56
PHOG	8	0,44	0,58	0,88	7,17
ReferenceColorSimilarity	10	0,47	0,53	1,01	**1124,32**
Tamura	7	0,44	0,59	0,80	6,42

Tabela 45 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento ROCK e o conjunto D_8. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	10	0,78	0,75	0,93	185,74
BIC	10	0,43	0,59	0,47	4,22
CEDD	10	0,44	0,59	1,05	1,17
FCTH	10	0,43	0,59	0,61	2,77
Gabor	10	0,43	0,42	0,39	2125,31
GCH	10	0,44	0,59	0,52	3,71
Haralick	10	0,46	0,51	0,40	**8150,98**
HaralickColor	10	0,46	0,54	0,68	4103,04
HaralickFull	14	0,43	0,58	0,26	19,61
JCD	10	0,43	0,59	0,67	2,43
LBP	10	0,43	0,59	0,76	22,62
LCH	10	0,44	0,59	1,05	0,93
Moments	10	0,46	0,52	0,65	4752,47
MPO	10	0,45	0,49	0,41	7668,71
MPOC	10	0,45	0,47	0,57	1379,66
PHOG	10	0,44	0,59	0,45	4,65
ReferenceColorSimilarity	10	**0,81**	**0,81**	0,45	1268,49
Tamura	10	0,44	0,59	**0,37**	7,89

APÊNDICE D – Resultados de todas as métricas obtidas para o conjunto D9

Tabela 46 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento AGNES e o conjunto D_9. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	8	0,55	0,56	1,26	722,37
BIC	8	0,50	0,53	0,91	810,45
CEDD	8	**0,74**	0,68	1,12	304,81
FCTH	8	0,68	0,68	1,26	1148,36
Gabor	8	0,35	0,34	0,53	5412,32
GCH	8	0,62	0,56	0,82	218,80
Haralick	8	0,46	0,48	0,52	9681,24
HaralickColor	8	0,50	0,53	0,63	3764,26
HaralickFull	8	0,60	0,56	**0,43**	3000,50
JCD	8	0,68	**0,68**	1,49	655,51
LBP	8	0,41	0,37	0,73	748,07
LCH	8	0,63	0,67	0,94	168,21
Moments	8	0,49	0,51	0,87	4387,37
MPO	8	0,44	0,48	0,48	**10755,56**
MPOC	8	0,40	0,43	0,82	2435,75
PHOG	8	0,53	0,52	1,42	48,48
ReferenceColorSimilarity	8	0,49	0,52	1,07	3446,50
Tamura	8	0,55	0,47	1,82	95,09

Tabela 47 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento CLARANS e o conjunto D_9. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	8	0,39	0,46	1,70	994,23
BIC	8	0,37	0,43	1,58	1078,64
CEDD	8	0,35	0,42	2,28	653,58
FCTH	8	0,40	0,47	1,40	1674,24
Gabor	8	0,29	0,33	0,57	9907,75
GCH	8	0,36	0,43	1,93	704,24
Haralick	8	0,36	0,43	**0,53**	15542,67
HaralickColor	8	0,39	0,44	0,87	6147,15
HaralickFull	8	0,37	0,40	0,58	8305,87
JCD	8	**0,52**	**0,60**	1,57	881,45
LBP	8	0,23	0,24	1,02	1789,19
LCH	8	0,39	0,50	1,79	407,87
Moments	8	0,40	0,45	0,95	8258,49
MPO	8	0,32	0,40	0,54	**16823,96**
MPOC	8	0,26	0,36	1,03	3956,91
PHOG	8	0,32	0,36	1,83	293,40
ReferenceColorSimilarity	8	0,46	0,50	1,03	4582,04
Tamura	8	0,39	0,43	1,94	336,38

Tabela 48 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento CURE e o conjunto D_9. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	8	0,80	0,78	0,82	264,25
BIC	8	0,48	0,47	0,90	196,17
CEDD	8	0,44	0,59	1,01	1,77
FCTH	8	0,78	0,75	0,66	719,56
Gabor	8	0,38	0,40	0,43	5999,03
GCH	8	0,64	0,57	0,71	143,77
Haralick	8	0,38	0,46	0,48	**17588,22**
HaralickColor	8	0,48	0,52	0,55	5175,71
HaralickFull	8	0,53	0,55	**0,41**	4444,63
JCD	8	0,43	0,59	0,97	3,09
LBP	8	0,38	0,37	0,61	738,45
LCH	8	0,44	0,59	0,54	3,88
Moments	8	0,47	0,52	0,53	6070,93
MPO	8	0,42	0,44	0,48	11083,37
MPOC	8	0,42	0,42	0,52	663,44
PHOG	8	0,44	0,59	0,65	6,13
ReferenceColorSimilarity	8	**0,81**	**0,81**	0,50	1602,17
Tamura	8	0,44	0,59	0,66	12,08

Tabela 49 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento DBSCAN e o conjunto D_9. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	6	0,58	0,54	1,15	94,45
BIC	8	0,43	0,59	0,96	4,18
CEDD	6	0,43	0,59	1,12	3,37
FCTH	6	0,43	0,59	1,38	4,12
Gabor	8	**0,49**	0,45	1,10	268,72
GCH	6	0,64	0,68	0,92	486,33
Haralick	6	0,45	0,56	1,24	14,78
HaralickColor	8	0,43	0,45	1,21	**665,11**
HaralickFull	8	0,49	0,53	0,93	54,55
JCD	7	0,43	0,59	1,00	3,91
LBP	6	0,44	0,59	0,60	15,15
LCH	8	0,43	**0,59**	1,11	2,93
Moments	7	0,41	0,48	2,14	133,42
MPO	8	0,47	0,54	1,07	36,06
MPOC	7	0,42	0,41	0,88	578,30
PHOG	8	0,44	0,59	**0,88**	7,17
ReferenceColorSimilarity	7	0,47	0,52	0,91	1429,50
Tamura	7	0,44	0,59	0,80	6,42

Tabela 50 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento FCM e o conjunto D_9. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	8	0,37	0,43	1,22	1262,47
BIC	8	0,33	0,41	1,26	1262,34
CEDD	8	0,37	0,45	1,98	718,97
FCTH	8	0,37	**0,47**	1,29	2036,56
Gabor	8	0,25	0,31	0,57	12941,16
GCH	8	0,33	0,44	2,21	744,09
Haralick	8	0,31	0,41	**0,54**	**22099,35**
HaralickColor	8	0,33	0,43	0,77	7819,19
HaralickFull	8	0,33	0,42	0,55	12021,33
JCD	8	0,35	0,43	1,89	1045,43
LBP	8	0,19	0,23	0,93	1958,94
LCH	4	0,52	0,59	1,45	576,96
Moments	8	0,32	0,41	0,82	10515,89
MPO	8	0,30	0,40	0,54	20527,57
MPOC	8	0,30	0,39	0,94	4657,58
PHOG	8	**0,39**	0,38	3,64	226,29
ReferenceColorSimilarity	8	0,33	0,43	1,00	5635,46
Tamura	8	0,31	0,39	3,46	331,18

Tabela 51 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento KMEANS e o conjunto D_9. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	8	0,37	0,44	1,18	1254,18
BIC	8	0,37	0,45	1,19	1252,58
CEDD	8	0,37	0,46	1,80	779,83
FCTH	8	**0,47**	**0,55**	1,27	1907,97
Gabor	8	0,26	0,32	0,56	12973,26
GCH	8	0,42	0,48	1,45	815,01
Haralick	8	0,33	0,43	**0,53**	**21803,48**
HaralickColor	8	0,37	0,46	0,72	7817,86
HaralickFull	8	0,35	0,43	0,53	11411,80
JCD	8	0,40	0,49	1,62	1082,25
LBP	8	0,20	0,24	0,91	1972,09
LCH	8	0,39	0,49	1,56	457,44
Moments	8	0,32	0,41	0,83	10455,94
MPO	8	0,31	0,40	0,54	20229,60
MPOC	8	0,32	0,40	0,91	4611,17
PHOG	8	0,29	0,33	1,77	319,40
ReferenceColorSimilarity	8	0,37	0,46	1,04	5197,58
Tamura	8	0,32	0,39	1,72	403,44

Tabela 52 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento KMEDOIDS e o conjunto D_9. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	8	**0,59**	**0,65**	1,25	589,19
BIC	8	0,42	0,51	1,29	982,95
CEDD	6	0,48	0,53	1,73	795,66
FCTH	6	0,48	0,52	1,74	1692,32
Gabor	8	0,25	0,31	0,58	12964,77
GCH	8	0,44	0,54	1,70	614,66
Haralick	8	0,32	0,41	0,59	18372,51
HaralickColor	8	0,41	0,48	0,71	6257,27
HaralickFull	8	0,32	0,39	**0,53**	10970,20
JCD	8	0,42	0,49	2,18	846,25
LBP	8	0,19	0,23	1,14	1603,74
LCH	6	0,62	0,68	2,01	278,88
Moments	8	0,34	0,42	0,82	10212,30
MPO	8	0,30	0,40	0,54	**20143,06**
MPOC	8	0,31	0,35	0,96	2593,92
PHOG	8	0,45	0,44	1,98	254,57
ReferenceColorSimilarity	8	0,45	0,52	1,05	4531,87
Tamura	7	0,54	0,52	2,19	254,91

Tabela 53 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento OPTICS e o conjunto D_9. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	6	0,58	0,54	1,15	93,89
BIC	8	0,43	0,59	0,96	4,18
CEDD	6	0,43	0,59	1,12	3,37
FCTH	6	0,43	0,59	1,28	3,45
Gabor	8	**0,50**	0,46	1,13	249,61
GCH	6	0,64	0,68	0,88	480,42
Haralick	6	0,45	0,57	1,25	13,24
HaralickColor	8	0,43	0,45	1,21	**659,42**
HaralickFull	8	0,49	0,53	0,94	50,44
JCD	7	0,43	0,59	1,00	3,91
LBP	6	0,44	0,59	0,60	15,15
LCH	8	0,43	**0,59**	1,11	2,93
Moments	7	0,42	0,50	2,05	130,71
MPO	8	0,47	0,55	1,09	31,14
MPOC	7	0,42	0,42	0,86	575,35
PHOG	8	0,44	0,59	**0,88**	7,17
ReferenceColorSimilarity	7	0,47	0,52	0,92	1390,86
Tamura	7	0,44	0,59	0,80	6,42

Tabela 54 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento ROCK e o conjunto D_9. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
ACC	8	0,78	0,75	1,23	238,77
BIC	8	0,44	0,59	0,47	4,34
CEDD	8	0,44	0,59	1,02	1,30
FCTH	8	0,43	0,59	0,61	2,76
GABOR	8	0,44	0,43	0,40	2624,93
GCH	8	0,44	0,59	0,51	3,93
HARACOLOR	8	0,47	0,56	0,37	5585,53
HARAFULL	8	0,47	0,55	0,65	4675,58
HARALICK	8	0,43	0,58	0,26	19,61
JCD	8	0,43	0,59	0,63	2,72
LBP	8	0,43	0,59	0,81	29,00
LCH	8	0,44	0,59	0,85	1,44
MOMENTS	8	0,47	0,56	0,60	4486,74
MPO	8	0,48	0,53	0,38	**6614,24**
MPOC	8	0,46	0,48	0,58	1607,25
PHOG	8	0,44	0,59	0,44	4,85
RCS	8	**0,81**	**0,81**	0,46	1629,90
TAMURA	8	0,44	0,59	**0,35**	8,64

APÊNDICE E – Resultados de todas as métricas obtidas para o conjunto D10

Tabela 55 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento AGNES e o conjunto D_{10}. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	4	0,45	0,57	1,34	606,17
BIC	4	0,45	0,52	0,73	903,32
CEDD	4	0,51	0,59	1,27	696,08
FCTH	4	0,56	0,60	1,15	1173,40
Gabor	4	0,39	0,51	**0,40**	3915,31
GCH	4	0,78	0,78	0,83	68,24
Haralick	4	0,49	0,51	0,47	7507,48
HaralickColor	4	0,33	0,43	0,53	5997,94
HaralickFull	4	0,61	0,63	0,42	2128,31
JCD	4	0,51	0,55	1,31	1174,81
LBP	4	0,69	0,70	0,54	550,41
LCH	4	0,65	0,67	1,25	109,52
Moments	4	0,35	0,43	0,49	8268,42
MPO	4	0,47	0,50	0,50	**8383,34**
MPOC	4	0,49	0,58	0,72	1667,49
PHOG	4	0,74	0,74	1,48	44,93
ReferenceColorSimilarity	4	**0,46**	**0,53**	1,28	4082,23
Tamura	4	**0,80**	**0,80**	**1,62**	62,23

Tabela 56 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento CLARANS e o conjunto D_{10}. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	4	0,29	0,42	1,12	1633,97
BIC	4	0,31	0,42	0,89	1666,73
CEDD	4	0,32	0,42	1,39	1273,21
FCTH	4	0,35	0,44	1,80	2482,44
Gabor	4	0,29	0,42	0,48	9336,72
GCH	4	0,30	0,42	1,07	1276,30
Haralick	4	0,33	0,42	**0,41**	**12852,79**
HaralickColor	4	0,30	0,42	0,49	7873,99
HaralickFull	4	0,43	0,47	0,56	7351,65
JCD	4	**0,43**	**0,50**	1,29	1606,47
LBP	4	0,33	0,43	0,98	2162,48
LCH	4	0,31	0,43	1,44	580,92
Moments	4	0,31	0,43	0,48	9633,83
MPO	4	0,34	0,43	0,43	10982,79
MPOC	4	0,30	0,42	0,70	5059,95
PHOG	4	0,35	0,44	1,79	446,74
ReferenceColorSimilarity	4	0,32	0,43	0,88	7134,94
Tamura	4	0,35	0,45	1,76	493,62

Tabela 57 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento CURE e o conjunto D_{10}. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	4	0,46	0,59	0,92	591,39
BIC	4	0,83	0,83	0,57	5,48
CEDD	4	**0,83**	**0,84**	0,92	1,48
FCTH	4	0,47	0,59	0,64	1657,15
Gabor	4	0,45	0,51	0,36	5606,59
GCH	4	0,61	0,64	0,77	321,62
Haralick	4	0,44	0,48	0,49	**9905,19**
HaralickColor	4	0,30	0,42	0,50	7562,97
HaralickFull	4	0,48	0,59	**0,34**	2056,90
JCD	4	0,83	0,83	0,97	1,92
LBP	4	0,75	0,75	0,41	394,91
LCH	4	0,83	0,84	0,82	6,29
Moments	4	0,45	0,51	0,49	4714,76
MPO	4	0,48	0,52	0,44	6708,73
MPOC	4	0,50	0,59	0,52	1521,53
PHOG	4	0,83	0,84	0,46	4,61
ReferenceColorSimilarity	4	0,48	0,59	0,66	3206,22
Tamura	4	0,83	0,83	0,55	15,23

Tabela 58 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento DBSCAN e o conjunto D_{10}. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	4	0,68	0,69	1,20	150,14
BIC	2	0,83	0,83	1,02	4,83
CEDD	2	0,83	0,83	1,35	2,70
FCTH	3	0,83	0,83	0,86	9,13
Gabor	3	0,81	0,81	1,84	20,23
GCH	4	0,44	0,51	1,02	782,44
Haralick	4	0,80	0,80	1,26	15,66
HaralickColor	4	0,66	0,66	1,97	288,44
HaralickFull	3	0,80	0,81	1,04	32,49
JCD	2	0,83	0,84	1,63	1,15
LBP	4	**0,83**	**0,83**	**0,69**	16,96
LCH	2	0,83	0,84	1,00	2,83
Moments	3	0,73	0,73	0,62	324,69
MPO	4	0,79	0,79	1,15	31,13
MPOC	4	0,53	0,60	1,89	874,26
PHOG	4	0,83	0,83	1,01	8,87
ReferenceColorSimilarity	4	0,39	0,44	1,08	**1413,72**
Tamura	2	0,83	0,83	0,61	13,36

Tabela 59 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento FCM e o conjunto D_{10}. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	4	0,30	0,42	1,00	1822,63
BIC	4	0,31	0,42	0,88	1735,35
CEDD	4	0,29	0,42	1,45	1277,09
FCTH	4	0,44	0,48	1,27	2761,02
Gabor	4	0,29	0,42	0,48	9600,82
GCH	4	0,29	0,42	1,06	1306,15
Haralick	4	0,31	0,42	**0,41**	**13527,54**
HaralickColor	4	0,30	0,42	0,48	7961,85
HaralickFull	4	**0,39**	0,45	0,57	7402,14
JCD	4	0,30	0,42	1,47	1769,58
LBP	4	0,31	0,43	0,98	2390,93
LCH	4	0,35	**0,48**	1,81	324,76
Moments	4	0,31	0,42	0,47	11115,41
MPO	4	0,30	0,42	0,47	12516,50
MPOC	4	0,34	0,44	0,69	5049,75
PHOG	4	0,35	0,44	2,51	391,65
ReferenceColorSimilarity	4	0,32	0,43	0,87	7187,25
Tamura	4	0,28	0,42	1,49	626,17

Tabela 60 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento KMEANS e o conjunto D_{10}. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	4	0,32	0,43	1,01	1726,92
BIC	4	0,33	0,43	0,90	1631,12
CEDD	4	0,38	0,45	1,43	1219,82
FCTH	4	**0,45**	**0,49**	1,19	2788,95
Gabor	4	0,30	0,42	0,48	9607,91
GCH	4	0,32	0,44	1,05	1254,85
Haralick	4	0,31	0,42	**0,40**	**13580,35**
HaralickColor	4	0,30	0,42	0,48	7966,81
HaralickFull	4	0,43	0,48	0,55	7405,99
JCD	4	0,42	0,47	1,42	1703,38
LBP	4	0,35	0,45	0,89	2320,99
LCH	4	0,36	0,45	1,40	560,06
Moments	4	0,31	0,42	0,47	11127,22
MPO	4	0,30	0,42	0,46	12545,93
MPOC	4	0,36	0,45	0,69	4968,35
PHOG	4	0,36	0,45	1,83	428,47
ReferenceColorSimilarity	4	0,38	0,45	0,92	6526,98
Tamura	4	0,27	0,42	1,45	634,75

Tabela 61 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento KMEDOIDS e o conjunto D_{10}. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	4	0,45	0,49	1,13	1274,27
BIC	4	0,44	0,49	1,02	1132,22
CEDD	4	0,32	0,43	1,29	1274,00
FCTH	4	0,32	0,43	1,39	2729,31
Gabor	4	0,30	0,42	0,48	9610,17
GCH	4	0,42	0,47	1,61	634,69
Haralick	4	0,31	0,42	**0,40**	**13579,51**
HaralickColor	4	0,30	0,42	0,47	7959,26
HaralickFull	4	0,36	0,44	0,57	7412,77
JCD	4	0,36	0,44	1,24	1624,16
LBP	4	0,31	0,43	1,00	2342,91
LCH	4	**0,46**	**0,50**	1,44	460,64
Moments	4	0,31	0,42	0,47	11126,55
MPO	4	0,30	0,42	0,46	12545,95
MPOC	4	0,40	0,47	0,69	4697,04
PHOG	4	0,42	0,47	1,68	442,42
ReferenceColorSimilarity	4	0,36	0,44	1,23	6037,77
Tamura	4	0,41	0,48	1,99	455,28

Tabela 62 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento OPTICS e o conjunto D_{10}. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
AutoColorCorrelogram	4	0,69	0,69	1,11	149,12
BIC	2	0,83	0,83	1,02	4,83
CEDD	2	0,83	0,83	1,35	2,70
FCTH	3	0,83	0,83	0,81	6,30
Gabor	3	0,81	0,81	1,85	19,46
GCH	4	0,45	0,51	1,00	778,58
Haralick	4	0,81	0,81	1,27	11,17
HaralickColor	4	0,67	0,67	1,98	286,06
HaralickFull	3	0,81	0,82	1,05	25,80
JCD	2	0,83	0,84	1,63	1,15
LBP	4	**0,83**	**0,83**	**0,63**	15,30
LCH	2	0,83	0,84	1,00	2,83
Moments	3	0,73	0,73	0,62	322,79
MPO	4	0,80	0,81	1,16	26,58
MPOC	4	0,54	0,60	1,86	863,78
PHOG	4	0,83	0,83	0,95	8,49
ReferenceColorSimilarity	4	0,39	0,44	1,08	**1397,78**
Tamura	2	0,83	0,84	0,55	9,45

Tabela 63 – Resultado das métricas (acurácia, FOWLKES, DAVIES e CALINSKI) obtidas por cada descritor e quantidade de grupos, considerando o algoritmo de agrupamento ROCK e o conjunto D_{10}. Em negrito são apresentados os melhores resultados (descritores) de acordo com cada uma das métricas.

DESCRITOR	Nº Grupos	ACURÁCIA	FOWLKES	DAVIES	CALINSKI
ACC	4	0,46	0,59	0,93	556,89
BIC	4	0,83	0,84	0,45	4,78
CEDD	4	0,83	0,83	0,74	1,76
FCTH	4	0,83	0,83	0,64	2,78
GABOR	4	0,67	0,68	0,34	1251,93
GCH	4	0,83	0,84	0,46	5,01
HARACOLOR	4	0,64	0,65	0,34	1107,21
HARAFULL	4	0,30	0,42	0,47	**7080,66**
HARALICK	4	0,82	0,82	0,26	19,61
JCD	4	0,83	0,83	0,55	3,19
LBP	4	0,82	0,83	0,37	67,47
LCH	4	**0,83**	**0,84**	0,94	1,18
MOMENTS	4	0,51	0,54	0,54	2663,56
MPO	4	0,83	0,83	0,41	15,28
MPOC	4	0,50	0,59	0,48	1332,25
PHOG	4	0,83	0,84	0,42	5,23
RCS	4	0,48	0,59	0,56	3197,07
TAMURA	4	0,83	0,84	**0,29**	12,21