Explorando Padrões Artísticos em Pinturas Impressionistas: Uma Abordagem de Aprendizado Não Supervisionado

Pedro Lucas Rangel Félix

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, PB
Junho - 2023

Pedro Lucas Rangel Félix

Explorando Padrões Artísticos em Pinturas Impressionistas: Uma Abordagem de Aprendizado Não Supervisionado

Monografia apresentada ao curso Ciência da Computação do Centro de Informática, da Universidade Federal da Paraíba, como requisito para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Thaís Gaudencio do Rêgo

João Pessoa, PB
Junho - 2023

F316e Félix, Pedro Lucas Rangel.
Explorando Padrões Artísticos em Pinturas Impressionistas: Uma Abordagem de Aprendizado Não Supervisionado / Pedro Lucas Rangel Félix. - João Pessoa, 2023.
58 f. : il.

Orientação: Thaís Gaudencio do Rêgo.
TCC (Graduação) - UFPB/CI.

1. Análise Computacional de Arte. 2. Inteligência Artificial. 3. Aprendizado Profundo. 4. Arte Impressionista. 5. Processamento de Imagem. I. Rêgo, Thaís Gaudencio do. II. Título.

UFPB/CI

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Trabalho de Conclusão de Curso de Ciência da Computação intitulado **Explorando Padrões Artísticos em Pinturas Impressionistas: Uma Abordagem de Aprendizado Não Supervisionado** de autoria de **Pedro Lucas Rangel Félix**, aprovada pela banca examinadora constituída pelos seguintes professores:

Thaís Gaudencio do Rêgo

Prof. Dr. Thaís Gaudencio do Rêgo
Universidade Federal da Paraíba

Yuri de Almeida Malheiros Barbosa

Prof. Dr. Yuri De Almeida Malheiros Barbosa
Universidade Federal da Paraíba

Tiago Maritan U. de Araújo

Prof. Dr. Tiago Maritan Ugulino de Araujo
Universidade Federal da Paraíba

João Pessoa, 22 de junho de 2023

Centro de Informática, Universidade Federal da Paraíba
Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil CEP: 58058-600
Fone: +55 (83) 3216 7093 / Fax: +55 (83) 3216 7117

“A dúvida é o princípio da sabedoria”
– Aristóteles

AGRADECIMENTOS

Não posso começar esta seção sem expressar minha profunda gratidão aos meus orientadores, a professora Thaís Gaudencio Do Rêgo e o professor Yuri De Almeida Malheiros Barbosa. A jornada acadêmica é um caminho repleto de desafios e incertezas, mas com vocês ao meu lado, encontrei força, inspiração e orientação para seguir em frente. Obrigado pela paciência incansável, pelo tempo investido e pela sabedoria compartilhada. O conhecimento e a compreensão que vocês me proporcionaram estenderam-se além dos limites desta pesquisa e permeiam todos os aspectos da minha vida. Vocês me inspiraram a buscar a excelência, a encarar os desafios com coragem e a nunca parar de aprender.

Agradeço também às minhas quatro companheiras felinas - Mia, Nyx, Catarina e Aparecida Carolina. Obrigado por estarem ao meu lado em cada momento desta jornada. Suas travessuras noturnas e corridas elétricas pela casa, embora tenham garantido algumas noites em claro, sempre conseguiram arrancar um sorriso do meu rosto, mesmo nos momentos mais estressantes. Seus miados constantes, que às vezes parecem nunca cessar, se tornaram uma trilha sonora reconfortante para meus longos dias de trabalho. Apesar dos momentos de caos felino, vocês me ensinaram a encontrar humor e alegria nos pequenos momentos e me lembraram da importância de tirar um tempo para apreciar as coisas simples da vida.

Finalmente, agradeço a todas as pessoas que, de uma forma ou de outra, contribuíram para a realização deste trabalho. Cada palavra de incentivo, cada gesto de apoio, cada momento compartilhado - tudo isso fez parte desta jornada. Este trabalho é o resultado não apenas do meu esforço, mas também do amor, apoio e orientação de todos vocês. Vocês são a força que impulsionou cada linha escrita, cada análise realizada, cada descoberta feita. Obrigado.

Dedico este trabalho aos meus avós maternos, cujo amor e carinho me fortaleceram durante esta jornada. Ao meu avô, cuja sabedoria transcende a mera acumulação de conhecimento. Sua interação respeitosa com a natureza, sua habilidade de ver poesia nos lugares mais simples, transformou profundamente minha percepção do mundo. Suas histórias e ensinamentos são os alicerces que embasam minha jornada acadêmica e pessoal.

Dedico também à minha querida avó paterna, que, embora não esteja mais entre nós, continua a ser uma presença constante e amorosa em minha vida. Suas memórias serão para sempre lembradas, como um farol que continua a me inspirar cada dia.

Este trabalho é uma expressão do meu amor e gratidão por todos vocês. Seu impacto em minha vida vai além das palavras e suas lições continuam a ressoar em minhas ações e decisões. Através deste trabalho, tento honrar seu legado, refletir sua sabedoria e expressar a profundidade do meu amor e respeito por todos vocês. Sua presença e orientação foram bússolas que me guiaram, e, embora este trabalho carregue o meu nome, é também um testamento do amor, da sabedoria e do carinho que vocês me proporcionaram.

RESUMO

Na confluência da arte e da ciência, o volume de pesquisas que exploram a aplicação de técnicas de inteligência artificial na análise de obras de arte tem crescido rapidamente. Este estudo buscou contribuir para esse campo em expansão, focando especificamente na análise computacional de pinturas impressionistas. Confrontados com a vastidão e a complexidade das obras artísticas, a tarefa de extrair conhecimento significativo representa um desafio computacional significativo. A implementação presente neste estudo segue com a redução de dimensionalidade das características extraídas de 304 pinturas impressionistas com PCA e t-SNE, e métodos de agrupamento, como *k-means* e agrupamento hierárquico. Utilizando o método do cotovelo, concluiu-se que cinco era o número ideal de agrupamentos, tanto para o PCA, quanto para o t-SNE. Esta descoberta foi reforçada pelo coeficiente de silhueta, que evidenciou uma coesão superior para o agrupamento *k-means* com $k = 5$. Os resultados também evidenciaram uma avaliação singular no que diz respeito ao gênero, com o método PCA e *k-means* com $k = 5$ apresentando uma entropia superior de gêneros, seja com 1,868 para dados não normalizados ou 2,341 para dados normalizados. Da mesma forma, esse método de agrupamento liderou a entropia de artistas, com um valor de 4,036 para dados não normalizados e 4,186 para dados normalizados. No entanto, no caso do t-SNE, embora o *k-means* com $k = 5$ também tenha demonstrado superioridade nos valores de entropia dos anos e artistas. Ainda assim, a métrica de entropia geral do t-SNE corroborou a eficácia do agrupamento *k-means* com $k = 5$.

Palavras-chave: <Análise Computacional de Arte>, <Inteligência Artificial>, <Aprendizado Profundo>, <Arte Impressionista>, <Processamento de Imagem>.

ABSTRACT

At the confluence of art and science, the volume of research exploring the application of artificial intelligence techniques in the analysis of works of art has grown rapidly. This study sought to contribute to this expanding field, focusing specifically on the computational analysis of impressionist paintings. Faced with the vastness and complexity of artistic works, the task of extracting meaningful knowledge represents a significant computational challenge. The implementation present in this study proceeded with the reduction of dimensionality of the features extracted from 304 impressionist paintings with PCA and t-SNE, and clustering methods such as k-means and hierarchical clustering. Using the elbow method, it was concluded that five was the ideal number of clusters, both for PCA and for t-SNE. This finding was reinforced by the silhouette coefficient, which evidenced superior cohesion for the k-means clustering with $k = 5$. The results also showed a unique evaluation with respect to genre, with the PCA method and k-means with $k = 5$ showing a higher genre entropy, whether with 1,868 for unnormalized data or 2,341 for normalized data. Similarly, this clustering method led the entropy of artists, with a value of 4,036 for unnormalized data and 4,186 for normalized data. However, in the case of t-SNE, although k-means with $k = 5$ also demonstrated superiority in the entropy values of years and artists. Still, the general entropy metric of t-SNE corroborated the efficacy of k-means clustering with $k = 5$.

Key-words: <Computational Art Analysis>, <Artificial Intelligence>, <Deep Learning>, <Impressionist Art>, <Image Processing>.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

CNN – Rede Neural Convolucional (do inglês, *Convolutional Neural Network*)

DBSCAN – Agrupamento Espacial Baseado em Densidade de Aplicativos com Ruído (do inglês, *Density-Based Spatial Clustering of Applications with Noise*)

DCEC – Agrupamento de Incorporação Convolucional Profundo (do inglês, *Deep Convolutional Embedding Clustering*)

HCA – Agrupamento Hierárquico Aglomerativo (do inglês, *Hierarchical Agglomerative Clustering*)

PCA – Análise de componentes principais (do inglês, *Principal Component Analysis*)

SVD – Decomposição em Valores Singulares (do inglês, *Singular Value Decomposition*)

t-SNE – Incorporação de vizinhança estocástica com distribuição t (do inglês, *t-Distributed Stochastic Neighbor Embedding*)

UFL – Aprendizado de Características não Supervisionado (do inglês, *Unsupervised Feature Learning*)

UFLK – Aprendizado de Característica não Supervisionados usando *k-means* (do inglês, *Unsupervised Feature Learning using k-means*)

VGG – Grupo de Geometria Visual *(*do inglês, *Visual Geometry Group)*

WSS – Soma dos quadrados intra-clusters (do inglês, *Within-Cluster-Sum of Squared Errors*)

SUMÁRIO

1. Introdução

A arte do Impressionismo, um movimento artístico que se originou na França no final do século XIX, é célebre pela sua tentativa de capturar a luz, a cor e a atmosfera de uma cena, em vez de se concentrar em detalhes precisos e linhas nítidas. Este movimento resultou em um conjunto de obras que, apesar de partilharem uma filosofia artística comum, apresentam uma diversidade considerável em termos de estilo, técnica e tema. Contudo, a análise destas obras de arte, devido à sua subjetividade intrínseca, pode ser uma tarefa complexa e, por vezes, elusiva.

Este trabalho é guiado pela hipótese de que a aprendizagem de máquina, em conjunto com técnicas de processamento de imagens, pode revelar padrões e tendências artísticas, identificar estilos específicos de artistas e examinar a evolução do movimento impressionista ao longo do tempo. A avaliação dos grupos é realizada em quatro setores principais: anos das pinturas, gêneros das pinturas, artistas e mediante uma métrica de avaliação matemática de agrupamentos que analisa a variância interna entre conjuntos.

Ao explorar a intersecção entre a arte e a ciência, este estudo oferece uma nova perspectiva para a análise de obras de arte, proporcionando novas formas de entender e apreciar o movimento impressionista.

1.1. Tema

Este estudo está ancorado na exploração do Impressionismo - um movimento artístico influente que revolucionou a pintura e escultura - através da lente da aprendizagem de máquina e do processamento de imagens. Utilizamos o potencial da ciência de dados e a capacidade de extração de recursos das redes neurais convolucionais para decifrar padrões complexos e sutilezas estilísticas presentes nas obras de arte impressionistas.

Ao focar no Impressionismo, buscamos não apenas entender as características estilísticas e temáticas únicas que definem este movimento, mas também explorar como a ciência de dados pode ser empregada para analisar e interpretar arte de forma inovadora. Este estudo serve como um passo significativo na junção de domínios - arte e ciência de dados - para gerar novas compreensões significativas e enriquecer nosso entendimento do movimento impressionista.

O tema deste trabalho sublinha o valor de abordagens interdisciplinares na pesquisa e estabelece um precedente para futuros estudos, que buscam explorar outras escolas artísticas, usando técnicas semelhantes de aprendizado de máquina e processamento de imagens.

1.2. Problema

A arte impressionista, com sua ênfase na captura de momentos efêmeros e as sutilezas de luz e cor, é notoriamente complexa e variada em sua expressão. A compreensão e análise dessas obras podem ser bastante desafiadoras, não apenas pela subjetividade inerente à interpretação da arte, mas também pela abundância de obras e artistas associados a este movimento.

Apesar da abundância de estudos e análises artísticas tradicionais sobre o impressionismo, existe uma lacuna significativa na aplicação de técnicas de aprendizado de máquina e processamento de imagens para a análise dessas obras de arte. As técnicas convencionais de análise de arte muitas vezes não conseguem captar totalmente a complexidade e a nuance presentes nas pinturas impressionistas. Além disso, a análise manual de grandes conjuntos de obras de arte pode ser um processo demorado e sujeito a viés.

O problema que este estudo busca resolver é, portanto, duplo: primeiro, como podemos aplicar efetivamente as técnicas de aprendizado de máquina e processamento de imagens para analisar e interpretar pinturas impressionistas? E segundo, como podemos fazer isso de maneira a revelar padrões que podem não ser facilmente perceptíveis através de métodos de análise tradicionais?

Este estudo é viável devido ao avanço recente nas técnicas de aprendizado de máquina e processamento de imagens, bem como a disponibilidade de grandes conjuntos de dados de pinturas impressionistas. No entanto, a complexidade e subjetividade da arte impressionista representam desafios significativos que devem ser abordados durante a execução deste projeto.

1.2.1 Objetivo geral

Nesse contexto, este trabalho propõe o uso de técnicas de aprendizado de máquina não supervisionado para a análise e agrupamento de pinturas impressionistas de modo a encontrar padrões e verificar qual o melhor método de agrupamento apresentado neste trabalho.

1.2.2 Objetivos específicos

Dentro do objetivo geral, os objetivos específicos podem ser estabelecidos por:

1. Pré-processamento e Extração de Características: Coleta e preparação de um conjunto de dados de pinturas impressionistas para análise, que envolve a seleção e redimensionamento das obras de arte e a extração de características visuais usando a rede pré-treinada VGG-19.

2. Redução de Dimensionalidade e Agrupamento: Aplicação de técnicas de redução de dimensionalidade, especificamente PCA e t-SNE, para simplificar a estrutura dos dados e facilitar a análise. Em seguida, a implementação do algoritmo k-means para agrupar as pinturas com características visuais semelhantes.

3. Avaliação e Comparação: Utilização do Coeficiente de Silhueta e análise dos atributos das pinturas para avaliar a qualidade dos grupos formados. Além disso, comparação dos resultados obtidos por diferentes técnicas de redução de dimensionalidade e agrupamento para determinar a abordagem mais eficaz para a análise de pinturas impressionistas.

1.3. Estrutura da monografia

No Capítulo 2, serão estabelecidas as bases teóricas necessárias para a compreensão das seções seguintes. Serão abordados tópicos como Transferência de Aprendizado, VGG-19 (Grupo de Geometria Visual, do inglês, Visual Geometry Group - VGG), Técnicas de Redução de Dimensionalidade, Análise de componentes principais (do inglês, Principal Component Analysis - PCA), Incorporação de vizinhança estocástica com distribuição t (do inglês, t-Distributed Stochastic Neighbor Embedding - t-SNE), *k-means*, Agrupamento Hierárquico, Coeficiente de Silhueta e Método do Cotovelo. Além disso, uma série de trabalhos relacionados no campo de aplicação será apresentada e comparada com as contribuições propostas neste estudo.

No Capítulo 3, serão explicitados todos os métodos utilizados para a elaboração da solução proposta. Esta seção incluirá detalhes sobre o conjunto de dados selecionado, a preparação desses dados e a arquitetura da solução proposta.

No Capítulo 4, será discutido o desempenho da metodologia proposta. Os resultados gerados serão analisados utilizando as métricas de avaliação estabelecidas, e as respectivas conclusões serão categorizadas e debatidas.

Finalmente, no Capítulo 5, serão apresentadas as conclusões do trabalho, fundamentadas no problema inicial, na solução proposta e nos resultados obtidos. Além disso, serão discutidas possíveis direções para trabalhos futuros no campo de estudo.

2. Fundamentação teórica

Nesta seção, definiremos os conceitos fundamentais indispensáveis para a compreensão das seções subsequentes, focando primordialmente em Transferência de Aprendizado, a estrutura e utilização do modelo VGG-19, e técnicas de redução de dimensionalidade, incluindo PCA e t-SNE. A discussão progredirá para a esfera de agrupamento de dados, abordando brevemente o algoritmo *k-means* e o método de Agrupamento Hierárquico, assim como o Coeficiente de Silhueta e o Método do Cotovelo.

2.1. Transferência de Aprendizado

A Transferência de Aprendizado é uma técnica de aprendizado de máquina que utiliza modelos pré-treinados em uma tarefa como ponto de partida para outra tarefa relacionada. Esta abordagem é particularmente útil em situações em que há escassez de dados disponíveis para treinamento ou quando o treinamento a partir do zero é computacionalmente proibitivo.

Os modelos de Aprendizado Profundo, como a VGG-19, são frequentemente treinados em grandes conjuntos de dados, como o ImageNet, que contém milhões de imagens em milhares de categorias distintas [1]. Estes modelos aprendem a identificar uma grande variedade de características, a partir dessas imagens, durante o processo de treinamento, desde características de baixo nível, como bordas e cores, até características de alto nível, como formas complexas e identidades de objetos.

Quando aplicamos a Transferência de Aprendizado, utilizamos os pesos aprendidos durante este processo de treinamento em uma nova tarefa. Em muitos casos fazemos um ajuste fino, ou seja, congelamos as camadas iniciais do modelo (ou seja, impedimos que seus pesos sejam alterados durante o treinamento) e treinamos apenas as camadas superiores do modelo na nova tarefa. Isto se baseia na ideia de que as características de baixo nível aprendidas pela rede são geralmente aplicáveis a uma ampla gama de tarefas, enquanto as características de alto nível são mais específicas

para a tarefa original [2]. Neste panorama, a rede VGG-19 se mostra como um excelente exemplo de um modelo pré-treinado que pode ser usado nesse tipo de abordagem.

2.2. VGG-19

A Rede Neural Convolucional do VGG, conforme delineada por Simonyan e Zisserman (2014), é uma entidade de referência na esfera da visão computacional, sendo reconhecida por sua arquitetura simples, porém eficiente. Essa rede neural compreende várias camadas convolucionais empilhadas, cada uma dotada de filtros de pequena extensão, seguidas por camadas plenamente conectadas, manifestando a lógica intrínseca da arquitetura VGG.

Dentre as múltiplas versões desta arquitetura, que variam conforme o número de camadas que compõem sua estrutura, a VGG-16 e a VGG-19 ganham destaque. Entretanto, para o propósito deste estudo, nos concentramos na versão VGG-19, composta por um total de 19 camadas, como podemos ver na Figura 1. Esta escolha se deve à sua capacidade aprimorada em comparação com as redes neurais convolucionais tradicionais, pois esta demonstra avanços na profundidade da rede. Sua estrutura alterna múltiplas camadas convolucionais e camadas de ativação não lineares, uma combinação que excede o desempenho de uma única camada convolucional [3] na extração de características de imagem ao detectar e discernir características mais complexas e específicas das imagens.

Em um nível mais aprofundado, a VGG-19 é caracterizada pelo uso exclusivo de filtros de convolução de tamanho 3x3, o menor tamanho capaz de capturar a noção de direção - esquerda/direita, cima/baixo, centro. Esta decisão arquitetural permite à VGG-19 alcançar campos receptivos equivalentes a redes com filtros maiores, mantendo a complexidade computacional e o número de parâmetros mais gerenciáveis.

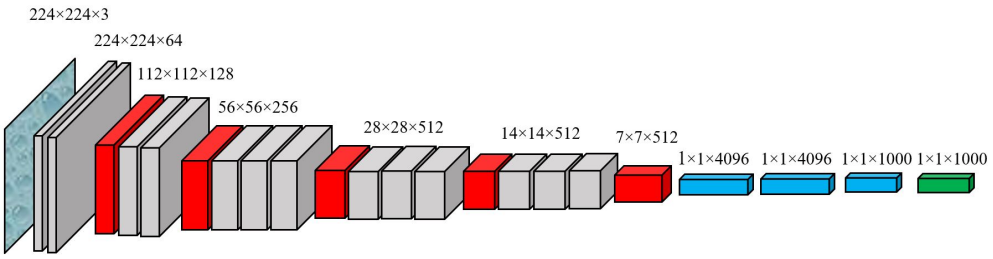

Figura 1: Esquema de camadas da rede VGG-19

Fonte: Wikimedia Commons

A VGG-19 também fez uma contribuição significativa para o campo do aprendizado profundo ao confirmar que a profundidade da rede - isto é, o número de camadas - é um fator crítico para um bom desempenho. No entanto, um efeito colateral disso é que ela acaba tendo um grande número de parâmetros (mais de 138 milhões) [4], tornando-a computacionalmente intensiva e desafiadora para treinar em termos de memória e poder de processamento. Esta necessidade de lidar com um grande número de parâmetros, como aqueles encontrados na VGG-19, nos conduz ao domínio das técnicas de redução de dimensionalidade.

2.3. TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE

Redução de dimensionalidade é uma técnica essencial em ciência de dados e aprendizado de máquina, usada para diminuir a complexidade computacional e melhorar a interpretabilidade de modelos complexos. Esta técnica é especialmente útil ao trabalhar com dados de alta dimensão, como imagens, onde cada *pixel* pode ser considerado uma dimensão distinta.

2.3.1. PCA

A PCA é uma técnica linear de redução de dimensionalidade, que busca identificar os componentes principais, ou direções de maior variância, em um conjunto de dados multidimensional. Uma vez identificados, esses componentes podem ser usados para projetar os dados originais em um espaço de menor dimensão, preservando o máximo de variância possível [5].

Em termos técnicos, a PCA realiza uma transformação ortogonal linear que converte um conjunto de variáveis correlacionadas, possivelmente, em um conjunto de variáveis não correlacionadas, chamadas componentes principais. O primeiro componente principal é o que tem a maior variância, o segundo componente principal é o que tem a segunda maior variância, e assim por diante. A ideia é reduzir a dimensão dos dados, ao mesmo tempo, em que preserva o máximo de informação (no sentido da variância) possível.

2.3.2. T-SNE

Diferente do PCA, o t-SNE é uma técnica não-linear de redução de dimensionalidade, particularmente bem adequada para a visualização de dados de alta dimensão. O t-SNE minimiza a divergência entre duas distribuições: uma distribuição que mede as semelhanças entre os pontos no espaço de alta dimensão e uma distribuição que mede as semelhanças entre os pontos no espaço de baixa dimensão [6].

O objetivo do t-SNE é encontrar uma representação de baixa dimensão [6] dos dados, que preserve o máximo possível as relações de vizinhança dos dados de alta dimensão. A técnica tem sido utilizada com sucesso na visualização de conjuntos de dados de grande dimensão, como aqueles encontrados em genômica e visão computacional.

Embora a PCA e o t-SNE sejam técnicas úteis para a visualização e manipulação de dados de alta dimensão, elas não nos fornecem informações explícitas

sobre a estrutura de agrupamento inerente aos dados. A fim de investigar possíveis agrupamentos, podemos recorrer a métodos específicos de agrupamento, como o algoritmo *k-means*.

2.4. ALGORITMO *K-MEANS*

O *k-means* (do português, *k*-médias) é um dos algoritmos de agrupamento mais comumente utilizados na aprendizagem de máquina. Esse algoritmo, introduzido por MacQueen em 1967, utiliza o conceito de centróides e visa minimizar a soma das distâncias entre cada ponto de dados e o centróide do conjunto ao qual foi atribuído [7].

Durante a inicialização, escolhe-se um número pré-determinado *k* de grupos. Os centróides de cada um desses segmentos são escolhidos de forma aleatória. Cada ponto de dados é atribuído ao agrupamento cujo centróide está mais próximo. A distância é geralmente calculada usando a distância euclidiana. Se *p* é um ponto de dados e *q* é o centróide, a distância euclidiana *d* entre eles é dada pela fórmula:

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

(Equação 2.4.1)

Após a atribuição de todos os pontos de dados, os centróides são recalculados tomando a média de todos os pontos de dados atribuídos a esse grupo. Os passos de cálculo e atribuição são repetidos até que uma condição de parada seja atendida. Geralmente, a condição de parada é quando os centróides não mudam significativamente entre duas iterações consecutivas, ou quando um número máximo de iterações é atingido.

2.5. AGRUPAMENTO HIERÁRQUICO

O Agrupamento Hierárquico é uma técnica poderosa e versátil de aprendizado não supervisionado. Nessa abordagem, construímos uma hierarquia de conjuntos que nos permite entender a estrutura subjacente dos dados [8]. A metodologia mais comum é a abordagem aglomerativa, também conhecida como Agrupamento Hierárquico Aglomerativo (do inglês, *Hierarchical Agglomerative Clustering* - HCA). No início do processo HCA, cada ponto de dados é tratado como um segmento individual. Se tivermos n pontos de dados, teremos N grupos no início.

Uma vez estabelecido esse ponto de partida, começamos a procurar os grupos mais próximos. A definição de "mais próximo" pode variar dependendo da medida de dissimilaridade que estamos usando. Normalmente medidas, como a distância euclidiana, são usadas para determinar a proximidade entre os conjuntos. Após calcular essas distâncias para todos os pares de grupos, identificamos os dois grupos mais próximos. Esses dois segmentos mais próximos são então aglomerados juntos, formando um único segmento. Este processo é conhecido como aglomeração. Com isso, reduzimos o número total de conjuntos por um.

Após a aglomeração, precisamos atualizar a matriz de proximidade para refletir a nova estrutura do grupo. Nesse ponto, o cálculo das distâncias deve ser refeito considerando a mudança na configuração dos agrupamentos. Este processo de encontrar os segmentos mais próximos, aglomerá-los juntos e atualizar a matriz de proximidade é repetido várias vezes. A cada iteração, o número total de conjuntos diminui em um. Continuamos este processo até que todos os pontos de dados sejam aglomerados em um único grupo.

Ao final desse processo, podemos visualizar o resultado como um dendrograma, uma representação gráfica, exemplificada pela Figura 2, que mostra a maneira pela qual os segmentos foram combinados.

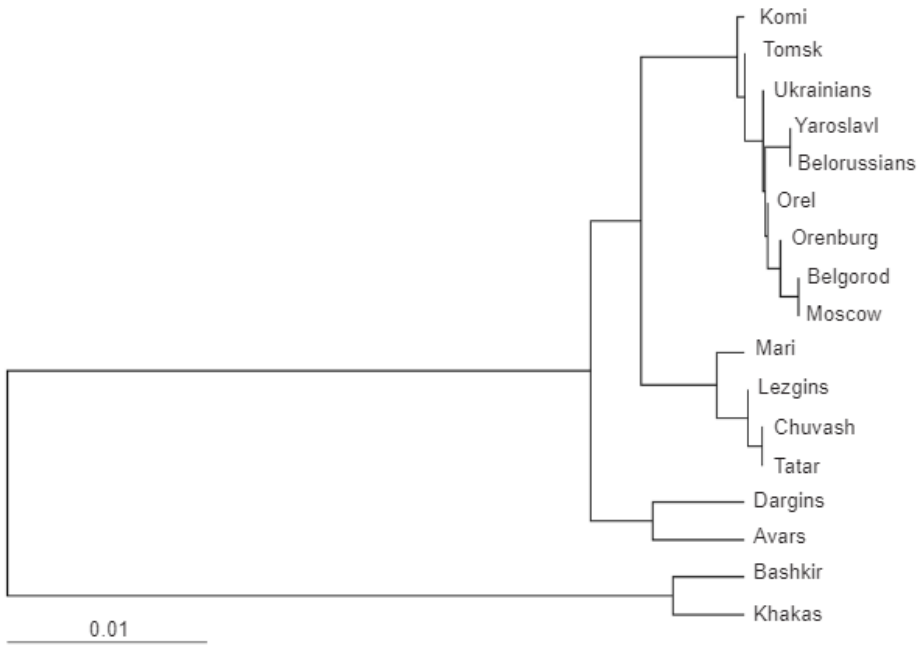

Figura 2: Exemplo de dendrograma das distâncias genéticas entre as populações russas

Fonte: Wikimedia Commons

O critério de ligação utilizado para calcular a proximidade entre os conjuntos é um aspecto crucial do agrupamento hierárquico. Algumas opções populares incluem a ligação simples, em que a distância entre dois grupos é definida como a distância entre os dois pontos mais próximos, um em cada grupo; a ligação completa, em que a distância é definida como a distância entre os dois pontos mais distantes; e a ligação média, em que a distância é a média das distâncias entre todos os pares de pontos, um de cada agrupamento.

2.6. Métricas de avaliação

A avaliação da qualidade de um agrupamento é uma etapa crucial na análise de conjunto. Duas métricas amplamente usadas para este fim são o Coeficiente de Silhueta e o Método do Cotovelo, detalhados a seguir.

2.6.1. Coeficiente de Silhueta

O Coeficiente de Silhueta é uma métrica que avalia a qualidade dos segmentos em uma análise de agrupamento. Ele fornece uma maneira quantitativa de avaliar a qual conjunto um determinado ponto de dados pertence mais provavelmente, ao comparar sua proximidade com os pontos dentro de seu próprio grupo e sua distância aos pontos em outros segmentos [9].

Dentro desta métrica, consideramos duas distâncias fundamentais. Primeiro, a distância média entre um ponto de dados e todos os outros pontos de dados em seu próprio conjunto é calculada. Essa medida é conhecida como distância intra-grupo. Este valor representa o quão bem o ponto de dados se encaixa dentro de seu próprio agrupamento. Por outro lado, também consideramos a distância inter-grupo, que é a média das distâncias entre um ponto de dados em um grupo e todos os pontos em outros grupos. Essa medida representa o grau de separação entre diferentes grupos de dados.

O Coeficiente de Silhueta é a diferença entre as distâncias inter-grupo e intra-grupo. O valor resultante vai de -1 a +1, onde um valor alto indica que um grupo de dados está coeso e um valor baixo sugere uma baixa coesão. A figura 3 exemplifica um gráfico de silhueta gerado a partir do agrupamento hierárquico com PCA, deste trabalho.

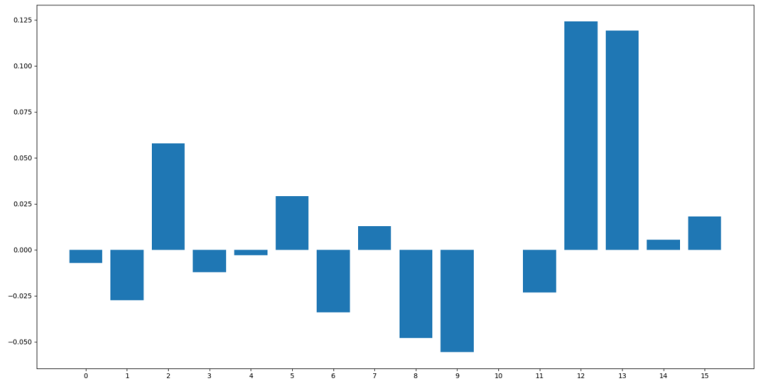

Figura 3: Exemplo de plotagem do método do coeficiente de silhueta, onde o eixo *x* representa os grupos e o eixo *y* representa os valores do coeficiente de silhueta

Fonte: autoria própria

2.6.2. Método do Cotovelo

O Método do Cotovelo, por outro lado, é usado principalmente para determinar o número ideal de segmentos em um algoritmo de agrupamento. Essa técnica envolve a execução do algoritmo de agrupamento para diferentes valores de k e o cálculo de uma medida de erro para cada k. Normalmente, essa medida de erro é a soma dos quadrados dentro do conjunto (ou seja, a soma das distâncias quadradas entre cada ponto de dados e o centróide do grupo ao qual pertence) [10].

Ao traçar essas medidas de erro em um gráfico com k no eixo x, a ideia é procurar um "cotovelo" na curva, um ponto onde a taxa de diminuição da medida de erro se torna significativamente menor, como visto na Figura 4, onde o eixo x representa a quantidade de grupos, o eixo y representa a soma das distâncias quadradas e a linha vertical representa o valor retornado pelo método. Esse ponto "cotovelo" é então escolhido como o valor ótimo para k.

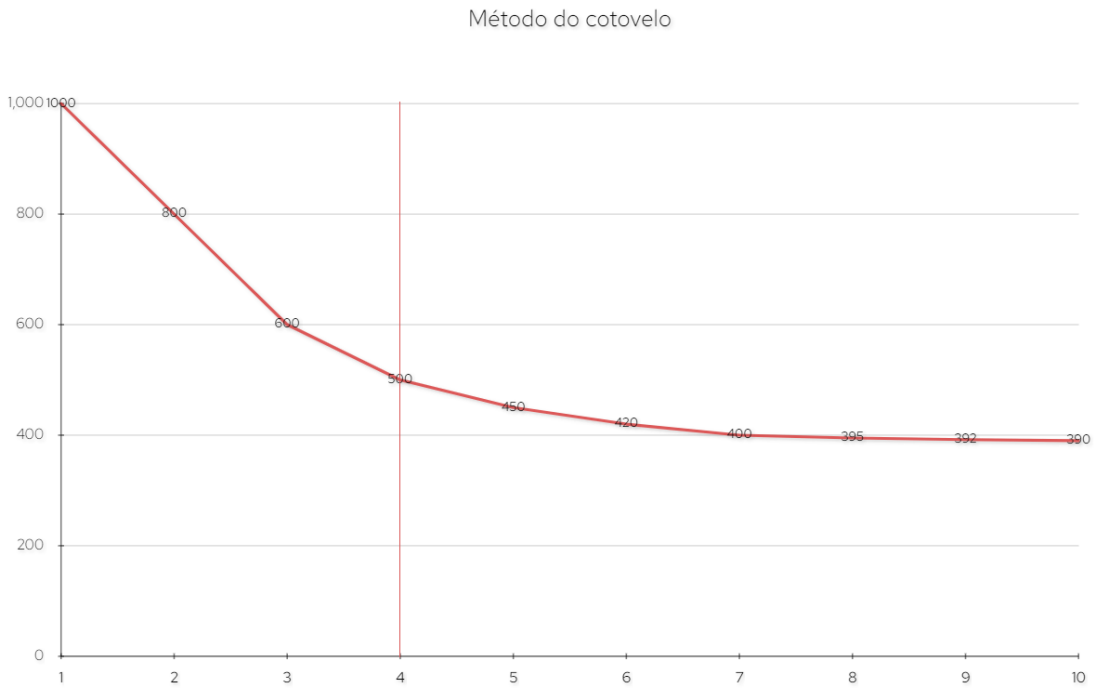

Figura 4: Exemplo de gráfico do método do cotovelo

Fonte: autoria própria

2.6.3. ENTROPIA DE SHANNON

A Entropia de Shannon é um conceito da teoria da informação, proposto por Claude E. Shannon em 1948, que mede a incerteza ou a impureza em um conjunto de dados [11]. A entropia quantifica a quantidade de informação esperada em um conjunto de dados, considerando a probabilidade de cada evento ocorrer. Matematicamente, a Entropia de Shannon para um sistema com N possíveis eventos é definida como:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad \text{(Equação 2.6.1)}$$

Aqui, p_i é a probabilidade de ocorrência do evento i. A base do logaritmo é 2 quando a informação é medida em bits [12].

Para entender o significado da fórmula, é útil considerar que p_i é a probabilidade de um evento, e $log2\ p_i$ é a quantidade de informação que obtemos quando o evento ocorre. Multiplicando esses dois fatores, obtemos a quantidade média de informação que esperamos obter quando o evento ocorre. Somamos todos esses produtos para todos os eventos para obter a entropia do sistema.

No contexto de agrupamento de dados, a Entropia de Shannon é usada para avaliar a qualidade dos agrupamentos. Um bom agrupamento seria aquele em que cada grupo contém elementos predominantemente de uma única classe. Em outras palavras, um grupo puro resultaria em menor incerteza e, portanto, menor entropia.

Grupos com baixa entropia indicam uma alta concentração de algumas classes específicas, sugerindo um agrupamento eficaz. Portanto, a entropia fornece uma ferramenta eficaz para avaliar e comparar diferentes métodos ou parâmetros de agrupamento [13].

2.7 TRABALHOS RELACIONADOS

Nesta seção, serão elucidados estudos relevantes que apresentam significativas contribuições em tópicos como agrupamento de imagens, Transferência de Aprendizado e redução de dimensionalidade. Estas publicações acadêmicas foram selecionadas a partir de uma pesquisa conduzida em plataformas acadêmicas, incluindo o Google Acadêmico, ACL *Anthology* e *Research Gate*. A estratégia de busca adotada para esta pesquisa utilizou termos-chave específicos como *impressionism* (impressionismo), *art* (arte), *image clustering* (agrupamento de imagens), *k-means*, VGG-19 e *agglomerative clustering* (Agrupamento Aglomerativo).

O trabalho [14] concentra-se na aplicação de aprendizado de características não supervisionado (do inglês, *Unsupervised Feature Learning* - UFL) para o reconhecimento de estilos artísticos. O trabalho apresenta uma nova abordagem para extrair características de pinturas e classificá-las. Os autores usaram um conjunto de dados de pinturas de diferentes estilos artísticos, como barroco, impressionismo, pós-impressionismo, realismo, *art nouveau*, romantismo, expressionismo e renascimento. Eles extraíram recursos dessas pinturas utilizando diferentes métodos, incluindo *pixels* brutos, PCA e Aprendizado de Característica não Supervisionados usando *k-means* (do inglês, *Unsupervised Feature Learning using k-means* - UFLK) com dimensões variadas.

Para cada estilo de pintura e método de extração de características, são fornecidas as métricas de desempenho (*F-score*, precisão e sensibilidade). O método com UFLK geralmente superou os métodos PCA e *pixels* brutos, em todos os estilos de pintura. Por exemplo, no caso do estilo barroco, o método UFLK com 3.000 dimensões alcançou um *F-score* de 0,644, significativamente maior do que o *F-score* obtido pelo método PCA com 1.000 dimensões (0,402) e o método de *pixels* brutos (0,412).

O artigo também inclui visualizações dos atributos a partir do UFLK e PCA para uma pintura intitulada *Scene from Tahitian Life* de Paul Gauguin. O UFLK apresenta linhas semelhantes, bordas e blocos sólidos de cores, e destacou elementos

específicos da pintura, como as saias das mulheres ou o fundo. Por outro lado, as feições do PCA foram mais parecidas com borrões, com pouca diferenciação de cores, e muitos detalhes da pintura foram perdidos na extração de feições do PCA.

Os autores também apresentam um gráfico de dispersão dos agrupamentos de estilos de pintura obtidos a partir do agrupamento espectral. O tamanho de cada nó, que representa uma única pintura no conjunto de dados, é determinado pelo seu grau (ou seja, o número de conexões que cada pintura tem com outras pinturas). A distância entre os nós demonstra a semelhança entre as pinturas. Pinturas e estilos semelhantes ou relacionados estão próximos, enquanto pinturas e estilos díspares estão distantes.

Em conclusão, o trabalho de pesquisa demonstra a eficácia do método UFLK para o reconhecimento de estilo artístico e fornece informações valiosas sobre as características de diferentes estilos de pintura.

O segundo trabalho [15] propõe um modelo de aprendizado profundo chamado DELIUS para agrupar artes visuais. O modelo usa recursos de abstração de alto nível extraídos de uma Rede Neural Convolucional (do inglês, *Convolutional Neural Network* - CNN) pré-treinada como uma entrada. O modelo não é forçado a aprender com as características para tarefas de classificação de estilo, mas explora de forma independente semelhanças visuais de alto nível entre as obras de arte para agrupá-las. Essa abordagem permite que o modelo leve em consideração características semânticas, relacionadas ao assunto e ao gênero da obra de arte, e não apenas propriedades estilísticas.

Os autores realizaram vários experimentos para avaliar a eficácia do método proposto. O primeiro experimento foi dedicado a avaliar a eficácia do método em agrupar os dados gerais. Um segundo experimento foi dedicado a mostrar a eficácia do método em agrupar obras de um mesmo artista. Em um terceiro experimento, os autores compararam a solução proposta com outras abordagens de agrupamento para o mesmo problema.

Os resultados dos experimentos mostraram que o modelo proposto foi capaz de agrupar obras de arte em grupos cada vez menores que compartilham semelhanças visuais. Curiosamente, os grupos encontrados tendem a refletir algumas influências

artísticas conhecidas e conexões entre artistas. Por exemplo, esboços e estudos de Duerer e Da Vinci, conhecidos por contribuírem significativamente para o renascimento, aparecem no mesmo grupo. Da mesma forma, o modelo tende a agrupar pinturas religiosas de artistas como Pontormo e van Eyck, bem como paisagens urbanas e rurais de Monet e Manet.

Os autores também experimentaram uma subamostra do conjunto de dados que compreende as obras de um único artista, Pablo Picasso. O modelo proposto mostrou uma tendência de agrupar obras com notáveis semelhanças visuais, relacionadas ao assunto e ao conteúdo da obra.

Em conclusão, o artigo apresenta uma nova abordagem para agrupar artes visuais usando aprendizado profundo. O modelo proposto, DELIUS, mostra resultados promissores no agrupamento de obras de arte com base em semelhanças visuais e características semânticas. No entanto, os autores reconhecem que há espaço para melhorias e mais pesquisas nessa área.

O terceiro trabalho [16] propõe um método para agrupar pinturas digitalizadas de maneira não supervisionada. Os autores apresentam um modelo chamado *DCEC-Paint*, que é baseado no modelo Agrupamento de Incorporação Convolucional Profundo (do inglês, *Deep Convolutional Embedding Clustering* - DCEC). O modelo é projetado para reconhecer padrões significativos em obras de arte visuais, uma tarefa tipicamente do domínio da percepção humana.

Os autores argumentam que reconhecer os atributos estilísticos e semânticos de uma pintura é extremamente difícil de conceituar, mas recursos relacionados ao visual, especialmente aqueles aprendidos por modelos de rede neural convolucional, podem ser eficazes na extração automática de padrões úteis de baixo nível, como características de cor e textura das obras de arte.

O modelo proposto foi testado em um conjunto de dados de pinturas de diferentes épocas e em uma subamostra composta apenas pelas obras de um único artista, Pablo Picasso. Os resultados mostraram que o modelo foi eficaz em encontrar grupos significativos em ambos os casos. O modelo também foi comparado com outras abordagens de agrupamento profundo, e superou-os.

Os autores sugerem que este modelo pode ser útil para muitas aplicações, incluindo ajudar especialistas em arte a encontrar tendências e influências entre escolas de pintura, descobrir diferentes períodos na produção do mesmo artista, descobrir quais obras de arte mais influenciaram o trabalho dos artistas atuais, apoiar a navegação interativa em galerias de arte, encontrando obras de arte visualmente vinculadas e ajudando os curadores a organizar melhor exposições permanentes ou temporárias com base em suas semelhanças visuais, em vez de motivações históricas.

Em trabalhos futuros, os autores planejam integrar informações contextuais que os usuários podem fornecer simplesmente olhando para uma obra de arte sem conhecimento prévio, a fim de tentar imitar a complexa percepção estética humana.

Na Tabela 1, está uma síntese dos trabalhos relacionados, apresentando informações sobre a base de dados utilizada, a arquitetura empregada e as métricas utilizadas. Também são mencionados os autores e uma breve descrição de seus objetivos.

Tabela 1: Resumo dos trabalhos relacionados

Autor	**Objetivo**	**Base de dados**	**Arquitetura**	**Métricas**
Eren Gultepe, Thomas E. Conturo e Masoud Makrehchi	Aplicação do UFL para o reconhecimento de estilos artísticos	Galeria de pinturas digitalizadas *abcgallery*	PCA, *k-means*, UFLK, Agrupamento Espectral	*F-score*, *F-mac*, sensibilidade, especificidade e precisão
Giovanna Castellano e Gennaro Vessio	Agrupamento de pinturas	*WikiArt*	*DenseNet121,* t-SNE	Coeficiente de silhueta, índice Calinski-Harabasz, precisão, informação mútua normalizada

Giovanna Castellano e Gennaro Vessio	Agrupamento de pinturas	*Best Artworks of All Time*, *Kaggle*	*DCEC-Paint,* t-SNE	Coeficiente de silhueta, índice Calinski-Harabasz
Este trabalho	Agrupamento de pinturas impressionistas seguindo múltiplas abordagens	*Painter by Numbers*	VGG-19, PCA, t-SNE, *k-means*, agrupamento hierárquico	Coeficiente de silhueta, entropia dos anos, artistas e gêneros

3. METODOLOGIA

Nesta seção, apresentamos a metodologia adotada para o agrupamento de pinturas impressionistas utilizando técnicas de aprendizado não supervisionado, como *k-means* com PCA e t-SNE e agrupamento hierárquico com PCA e t-SNE. Descrevemos os dados utilizados, o pré-processamento das imagens, a extração de características com a rede VGG-19, a redução de dimensionalidade e as técnicas de agrupamento.

3.1. FERRAMENTAS

Para a implementação proposta, a linguagem de programação *Python* (versão 3.7.4) com o apoio das seguintes bibliotecas: *Numpy* (versão 1.23.4) para o auxílio da gestão de estruturas complexas de dados, *Pandas* (versão 1.5.1) para a leitura e criação de planilhas, *Keras* (versão 2.9.0) para a aplicação do modelo VGG-19 e *Scikit-learn* (versão 1.1.3) para a realização dos agrupamentos e aplicação do coeficiente de silhueta.

O sistema foi montado usando um computador equipado com 16GB de memória RAM, composta por dois módulos de 8GB, operando a uma frequência de 3600MHz. Além disso, possui um SSD de 1TB para armazenamento, um processador Intel(R) Core™ i5-9300H rodando a 2.40GHz e uma placa de vídeo NVidia GTX 1650 com 4GB de memória dedicada.

3.2. DADOS

O pré-processamento das imagens é um passo crítico para garantir que os algoritmos de aprendizado não supervisionado possam extrair padrões significativos e representativos das pinturas. A base de dados utilizada é a *Painter by Numbers*, fornecida pela empresa *Kaggle*, a qual é separada em dados de treino e dados de teste.

Utilizamos as imagens de treino, onde se somam 23.817 pinturas de técnicas artísticas diversas.

Com base no banco de dados proposto, foi construída uma amostra, como podemos ver alguns exemplos na Figura 5, para seguir com a proposta final deste trabalho. As imagens foram filtradas para só serem adicionadas em nosso banco de dados final se forem do estilo artístico impressionismo e se contiverem o ano catalogado, utilizado em forma de análise a seguir, totalizando 304 imagens.

Figura 5: Exemplo de imagens contidas na amostra final

Fonte: autoria própria

3.2.1. Processamento das imagens

As imagens foram pré-processadas para garantir a consistência e facilitar a análise. Todas as imagens foram redimensionadas para 256x256 *pixels*. Além disso, houve a necessidade de realização de um pós-processamento manual de remoção de

certos elementos das pinturas, pois algumas das imagens adicionadas em nosso banco de dados final haviam moldura, como mostra a Figura 6, e pequenas palavras indicando sua fonte digitalizadora, aspectos que não contribuem com a análise proposta neste trabalho. Com os ajustes, a extração das características visuais das imagens se deu pelo modelo pré-treinado VGG-19.

Figura 6: Exemplo de imagem com moldura

Fonte: autoria própria

3.3. REDUÇÃO DE DIMENSIONALIDADE

As imagens de pinturas e as características extraídas por redes neurais convolucionais, como a VGG-19, possuem um vetor de saída de dimensionalidade igual a 4096, o que pode dificultar a análise e interpretação dos dados. A redução de dimensionalidade é uma etapa crucial para simplificar a estrutura dos dados, facilitar a visualização e melhorar a eficiência dos algoritmos de aprendizado não supervisionado. Nesta seção, investigamos a aplicação de duas técnicas populares de redução de dimensionalidade - PCA e t-SNE.

3.3.1. PCA

O PCA foi aplicado para reduzir a dimensionalidade dos dados das imagens, preservando a maior variância possível. A biblioteca utilizada foi o *sklearn,* onde do módulo *decomposition* foi importada e instanciada a classe PCA onde a quantidade de componentes principais, representada pela propriedade *n_components*, foi estabelecida aleatoriamente em 100. Essa quantidade determina o número de componentes principais a serem mantidos após a redução de dimensionalidade. Da mesma forma, o parâmetro *random_state*, que controla a aleatoriedade do algoritmo, foi fixado arbitrariamente como 22 para garantir a reprodutibilidade dos resultados, embora este seja um valor escolhido sem uma justificativa específica. Essa abordagem de seleção aleatória para os parâmetros foi adotada para avaliar a robustez do algoritmo perante diferentes cenários. Após isso, foi passado como parâmetro de ajuste e de transformação o *array* com as características extraídas das pinturas.

3.3.2. T-SNE

O t-SNE foi utilizado como outra técnica de redução de dimensionalidade, que preserva as relações de distâncias locais entre as imagens. A biblioteca utilizada foi o *sklearn*, onde do módulo *manifold* foi importada a classe t-SNE e utilizada com a parametrização padrão da biblioteca, gerando um vetor de dimensionalidade igual a 2. Após isso, foi passado como parâmetro de ajuste e transformação o *array* com as características extraídas das pinturas.

3.4. AGRUPAMENTO

Ao aplicar técnicas de agrupamento em pinturas impressionistas, podemos revelar padrões e tendências artísticas, como a semelhança visual entre pinturas, identificar estilos específicos de artistas e examinar a evolução do movimento ao

longo do tempo. Nesta seção, abordamos o uso de dois algoritmos de agrupamento amplamente utilizados - *k-means* e agrupamento hierárquico.

3.4.1. ALGORITMO *K-MEANS*

O algoritmo *k-means* foi aplicado para agrupar as pinturas, utilizando as representações de dimensão reduzida obtidas com o PCA e t-SNE. O algoritmo *k-means* foi aplicado para agrupar as pinturas, utilizando as representações de dimensão reduzida obtidas com o PCA e t-SNE. Para calcular um número otimizado de conjuntos, o teste do cotovelo foi usado. Utilizando-se da biblioteca *sklearn*, mais especificamente do módulo *cluster*, foi importada a classe *k-means* e instanciada com *random_state* = 22 e o que variou foi a quantidade de grupos, representada pelo parâmetro *n_clusters,* que assumiu os valores de 2 a 21, valores estes arbitrários para a realização do teste do cotovelo. A partir das opções de agrupamento geradas, a biblioteca *kneed* foi utilizada para encontrar programaticamente o cotovelo, usando-se de uma implementação proposta por Ville Satopaa et al. (2011).

A partir do Método do Cotovelo, o número otimizado de segmentos foi 5, tanto para o método que utiliza PCA, quanto para o que utiliza o t-SNE. Para a realização de comparações acerca dos métodos de agrupamento, a partir dos resultados do agrupamento hierárquico foram geradas para o PCA e t-SNE novas opções de agrupamentos, sendo essas opções $k = 8$ e $k = 16$.

3.4.2. AGRUPAMENTO HIERÁRQUICO

O Agrupamento Hierárquico aglomerativo foi utilizado como outra abordagem de agrupamento, empregando as representações de dimensão reduzida obtidas através do PCA e do t-SNE. A implementação foi realizada com o uso da classe *AgglomerativeClustering*, proveniente do módulo de agrupamento da biblioteca *sklearn*. Os parâmetros selecionados de forma aleatória foram *distance_threshold* definido como 120 e *n_clusters* como *None* para o método PCA, enquanto para o

t-SNE foram utilizados os parâmetros *distance_threshold* definidos como 20 e n_clusters como None.

O valor de *distance_threshold* controla o nível de similaridade necessário para fundir conjuntos durante o processo aglomerativo, ou seja, os grupos são aglomerados até que a distância entre eles exceda esse limiar. A escolha desses valores não foi motivada por uma análise prévia dos dados, mas de maneira arbitrária, com o objetivo de explorar a eficácia do algoritmo quando operado nessas condições. Por outro lado, o parâmetro *n_clusters* foi definido como *None*, significando que o número de grupos seria determinado automaticamente pelo algoritmo com base no limiar de distância estipulado.

A visualização dos resultados foi realizada com a criação de um dendrograma, gerado com o auxílio da biblioteca *matplotlib* e a função *dendrogram*, presente na biblioteca *scipy*. No entanto, isso foi realizado apenas para os resultados obtidos com o PCA, devido à natureza mutável do t-SNE que não permite replicar os mesmos padrões, uma vez que o código é executado novamente. Com o PCA, obteve-se um total de 16 grupos, enquanto com o t-SNE, o número de grupos foi 8.

3.5. AVALIAÇÃO DOS GRUPOS

Para analisar os grupos obtidos com cada técnica, foram gerados gráficos de dispersão das representações bidimensionais dos dados, com pontos coloridos conforme os segmentos atribuídos pelo algoritmo *k-means* e pelo agrupamento hierárquico. Essas visualizações facilitaram a comparação das técnicas e a identificação de padrões e agrupamentos nos dados. Os conjuntos obtidos também foram avaliados em quatro categorias principais: anos das pinturas, gêneros das pinturas, artistas e coeficiente de silhueta. Para as categorias anos, gêneros e artistas, os dados foram analisados de forma normalizada e sem estarem normalizados, a fim de explorar possíveis diferenças nos resultados. Na Tabela 2 conseguimos ver quais membros estão integrando os critérios de análise: gênero, artista e ano.

Tabela 2: Lista de gêneros e artistas presentes em nossa base de dados

Critério	Lista de membros
Gênero	Pintura de gênero (*genre painting*), paisagem com nuvens (*cloudscape*), paisagem (*landscape*), esboço e estudo (*sketch and study*), autorretrato (*self-portrait*), pintura de flores (*flower painting*), pintura religiosas (*religious painting*), retrato (*portrait*), pintura de batalha (*battle painting*), pintura de animal (*animal painting*), natureza-morta (*still life*), pintura do nu artístico (*nude painting*), marinha (*marina*), paisagem de cidade (*cityscape*), interior (*interior*), ilustração (*illustration*)
Artista	Zinaida Serebriakova, Berthe Morisot, At Les Ambassadeurs, Edgar Degas, James McNeill Whistler, Gustave Caillebotte, Arthur Streeton, Iosif Iser, Claude Monet, Vilhelm Hammershoi, John Henry Twachtman, Tom Roberts, Max Liebermann, John Singer Sargent, Francis Picabia, Camille Pissarro, Mihaly Munkacsy, Ivan Grohar, Giovanni Fattori, Paul Gauguin, Philip Wilson Steer, Howard Pyle, Konstantin Yuon, Akseli Gallen-Kallela, James Tissot, T. C. Steele, John Lavery, Stefan Dimitrescu, Robert Julian Onderdonk, Pierre-Auguste Renoir, Konstantin Korovin, Henry Herbert La Thangue, Henri Matisse, Valentin Serov, Childe Hassam, Guy Rose, Pablo Picasso, Theo van Rysselberghe, Filipp Malyavin, Egon Schiele, Arkhip Kuindzhi, Isaac Levitan, Georges Seurat, Alfred Sisley, Maurice Prendergast, Thomas Pollock Anshutz, John Marin, Eva Gonzales, Eugene Boudin, William Merritt Chase, Henri de Toulouse-Lautrec, Henry Ossawa Tanner, Willard Metcalf, Vasile Popescu, Julian Alden Weir, Mary Cassatt, Federico

	Zandomeneghi, Carlo Carra, William James Glackens, Edouard Manet, Paul Cezanne
Ano	1861, 1865, 1866, 1870, 1871, 1872, 1873, 1874, 1875, 1876, 1877, 1878, 1879, 1880, 1881, 1882, 1883, 1884, 1885, 1886, 1887, 1888, 1889, 1890, 1891, 1892, 1893, 1894, 1895, 1896, 1897, 1898, 1899, 1900, 1901, 1902, 1903, 1904, 1905, 1906, 1907, 1908, 1909, 1910, 1912, 1913, 1914, 1915, 1916, 1917, 1918, 1919, 1920, 1921, 1923, 1926, 1929, 1930, 1937

Fonte: autoria própria

3.5.1. Anos das pinturas

Para a análise temporal dos grupos, foram gerados gráficos de violino representando a distribuição dos anos das pinturas. Esses gráficos permitiram a identificação de possíveis agrupamentos temporais e a comparação das técnicas em relação a essa variável. A entropia foi utilizada como uma medida da desordem ou aleatoriedade em cada agrupamento. Especificamente, mede a diversidade ou a dispersão dos itens em um determinado grupo. Os agrupamentos com maior entropia média sugerem uma não homogeneidade dos dados, mostrando assim que há picos de concentração dos dados em determinados grupos.

3.5.2. Gêneros das pinturas

A dispersão dos gêneros das pinturas nos agrupamentos foi analisada por meio de gráficos de barra, tanto para a versão padronizada, quanto para a versão sem padronização. Esta avaliação possibilitou medir a eficácia das técnicas na identificação de agrupamentos temáticos e estilísticos presentes nas pinturas. Na análise dos gêneros das pinturas, também utilizamos a entropia como método avaliativo

3.5.3. Artistas

A disposição dos artistas nos segmentos também foi estudada utilizando gráficos de barra, tanto na versão padronizada, quanto na não padronizada. Este estudo permitiu discernir se os grupos representam estilos particulares de artistas ou se reúnem pinturas de diferentes artistas com atributos visuais similares. No contexto da análise de artistas, o valor da entropia foi igualmente aplicado, mas desta vez considerando a distribuição de obras de arte por artista dentro de cada agrupamento.

3.5.4. Coeficiente de Silhueta

O Coeficiente de Silhueta foi calculado para avaliar a qualidade dos conjuntos, a partir das representações de dimensão reduzida com PCA. Devido à natureza não determinística do t-SNE, o gráfico de silhueta só foi possível realizar para o PCA, uma vez que ao rodar novamente o código com t-SNE, grupos diferentes são gerados. Foi utilizada a função *silhouette_samples*, provinda do módulo de métricas do *sklearn* e foram passadas como parâmetro as características extraídas com a dimensionalidade reduzida e o valor resultante do eixo *y* proveniente do *k-means*. Para cada agrupamento foi gerado um valor de silhueta, representados em um gráfico utilizando a biblioteca *matplotlib*.

4. Apresentação e análise dos resultados

Nesta seção, iremos analisar e discutir os resultados obtidos a partir da aplicação da metodologia, descrita na seção anterior, ou seja, apresentaremos os achados decorrentes do agrupamento de pinturas impressionistas. Utilizou-se, para tanto, técnicas de aprendizado não supervisionado, com ênfase no *k-means* e no agrupamento hierárquico em conjunto com PCA e t-SNE. A seção foi estruturada em duas partes: na primeira, apresentaremos os resultados com base em métricas de avaliação; e na segunda faremos uma avaliação comparativa entre os métodos de agrupamento propostos neste estudo.

4.1. Métricas de avaliação

Nesta seção procuraremos entender as relações presentes nos dados através das métricas selecionadas: ano, gênero, artista e coeficiente de silhueta.

4.1.1 PCA

A partir do método pré-avaliativo, temos que o método do cotovelo [14] nos retorna um valor idealizado para o número de agrupamentos, mostrado na Figura 7, onde o eixo x nos retorna a quantidade de grupos avaliada e o eixo y nos retorna o valor da soma da distância ao quadrado entre cada ponto e o centróide do grupo. Para o PCA esta métrica retorna como valor final cinco.

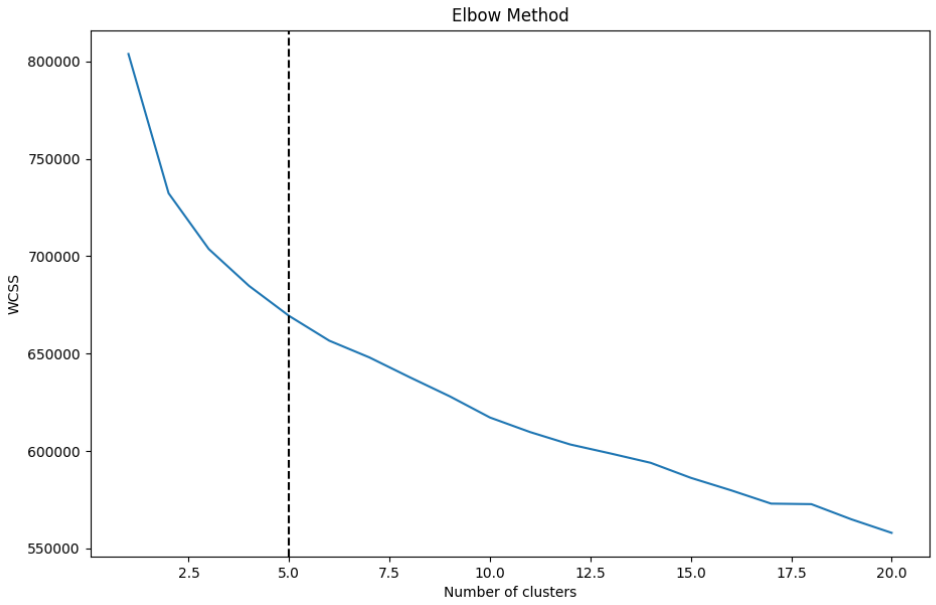

Figura 7: Método do cotovelo para o PCA

Fonte: autoria própria

Em comunhão com os dados apresentados pelo método do cotovelo, através do coeficiente de silhueta conseguimos estender esta análise e avaliar a média da dispersão geral de cada método de agrupamento, assim como mostra a Tabela 3. A partir do maior valor médio conseguimos ter como resultado o agrupamento com a maior coesão em comparação com os outros métodos.

Tabela 3: Média do coeficiente de silhueta para cada método de agrupamento que utiliza PCA

Métrica de avaliação	***k* = 5**	***k* = 8**	***k* = 16**	**Agrupamento hierárquico**
Média do coeficiente de silhueta	0,042	0,017	0,017	0,010

Fonte: autoria própria

Partindo da maior média do coeficiente de silhueta, conseguimos encontrar o agrupamento mais coeso, que neste caso seria o do método *k-means*, com $k = 5$. Seguindo com a avaliação por ano, notavelmente, a entropia que, com $k = 5$ evidencia um valor superior, registrando 4,903. Isso contrasta com os outros métodos analisados: PCA Hierárquico (3,294), PCA com $k = 16$ (3,266) e PCA com $k = 8$ (4,5). Portanto, conforme o critério da entropia, o método PCA com $k = 5$ supera os demais.

A próxima etapa de nossa análise consiste em identificar o melhor agrupamento da perspectiva de gêneros, levando em conta a tabela com os dados normalizados e não normalizados.

Tabela 4: Entropia média dos gêneros

Métrica de avaliação	***k-means* com $k = 5$**	***k-means* com $k = 8$**	***k-means* com $k = 16$**	**Agrupamento hierárquico**
Entropia dos gêneros com dados não normalizados	1,868	1,702	1,304	1,425
Entropia dos gêneros com dados normalizados	2,341	1,949	1,178	1,374

Fonte: autoria própria

A métrica de gêneros nos revela, a partir da perspectiva da entropia, que o agrupamento com média superior aos outros seria o método que utiliza o *k-means*, com $k = 5$, com 2,341 para os dados não normalizados e 1,868 para os dados normalizados.

Como última métrica estabelecida, temos os artistas. Conseguimos traçar uma coesão a partir da entropia também, como vemos na Tabela 5, e então verificar o maior valor para a nossa análise.

Tabela 5: Entropia média dos artistas

Métrica de avaliação	***k-means* com *k* = 5**	***k-means* com *k* = 8**	***k-means* com *k* = 16**	**Agrupamento hierárquico**
Entropia dos artistas com dados não normalizados	4,036	3,725	2,773	2,783
Entropia dos artistas com dados normalizados	4,186	3,655	2,627	2,556

Fonte: autoria própria

Pela métrica de entropia, conseguimos ver que o método que utiliza o *k-means* com $k = 5$ novamente obteve melhores resultados, em relação aos outros métodos de agrupamento com PCA utilizados comparativamente neste trabalho. A partir do método do cotovelo, onde o resultado foi 5 para a quantidade de grupos ideal, as outras métricas de avaliação (anos, artistas e gêneros) seguem o comportamento da quantidade ideal de grupos como 5 também, sugerindo que para o método PCA o melhor agrupamento seguiu o valor resultante a partir do método do cotovelo.

Para o complemento da nossa análise, conseguimos adotar o coeficiente de silhueta localmente em um agrupamento. Tomando como base o melhor método resultado da nossa análise prévia, o com *k-means* e $k = 5$, conseguimos traçar os dados, como visto na Tabela 6.

Tabela 6: Coeficientes de silhueta para cada grupo, com *k-means* e *k* = 5

Métrica de avaliação	Grupo 0	Grupo 1	Grupo 3	Grupo 4	Grupo 5
Coeficiente de silhueta	0,140	0,018	-0,004	0,011	-0,020

Fonte: autoria própria

A partir dos coeficientes de silhueta, temos que o grupo mais coeso dentro do método com *k-means* e *k* = 5 seria o primeiro grupo. A fins demonstrativos conseguimos ver um exemplo de imagens presentes neste grupo, como vemos na Figura 8.

Figura 8: Exemplo do primeiro grupo gerado a partir do método utilizando PCA com *k-means* e *k* = 5

Fonte: autoria própria

4.1.2. t-SNE

A partir do método pré-avaliativo, temos que o método do cotovelo [17] também nos retorna o valor igual à cinco, mostrado na Figura 9.

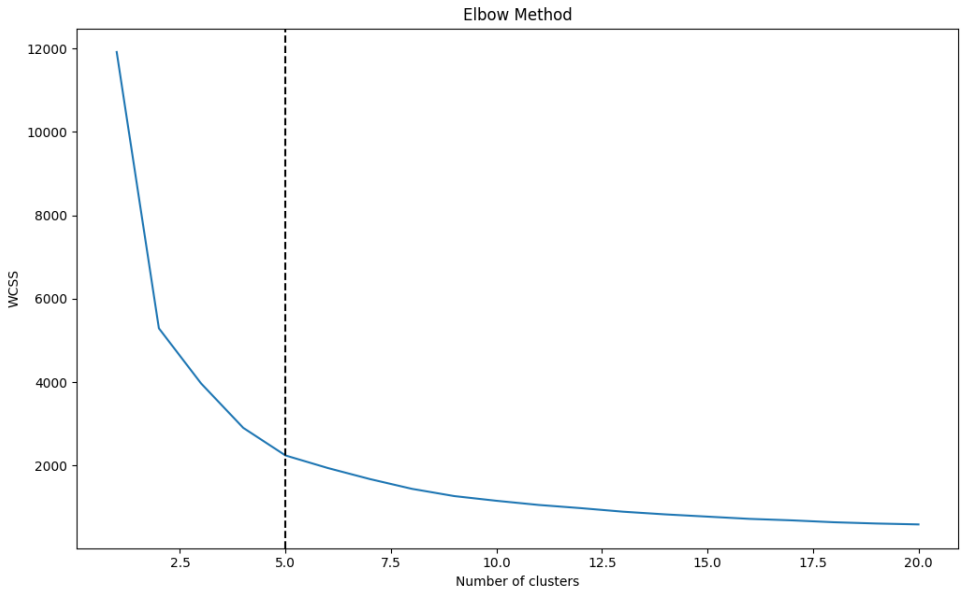

Figura 9: Método do cotovelo para o t-SNE

Fonte: autoria própria

Seguindo os passos avaliativos, assim como para o método PCA, não conseguimos os valores das métricas para o coeficiente de silhueta para o t-SNE, devido à sua característica não determinística, ou seja, uma vez que o código é executado não conseguimos replicar os valores novamente.

Seguindo com a análise de anos, o método t-SNE com $k = 5$ apresenta melhores resultados, com uma entropia de 4,925. Este é seguido por $k = 8$ (4,517), t-SNE Hierárquico (4,451), e $k = 16$ (3,825). Assim, pelo critério da entropia dos anos, o método utilizando *k-means* com $k = 5$ demonstra ser o mais eficaz entre os métodos t-SNE examinados.

Para as métricas de gêneros e artistas, podemos seguir o mesmo método utilizado para a análise do método com PCA. A partir da entropia de gêneros e artistas podemos traçar os resultados, como podemos ver na Tabela 7.

Tabela 7: Entropia média dos gêneros e artistas

Métrica de avaliação	***k-means* com *k* = 5**	***k-means* com *k* = 8**	***k-means* com *k* = 16**	**Agrupamento hierárquico**
Entropia dos gêneros com dados não normalizados	1,877	1,719	1,435	1,674
Entropia dos gêneros com dados normalizados	2,374	1,864	1,351	1,935
Entropia dos artistas com dados não normalizados	4,138	3,845	3,198	3,836
Entropia dos artistas com dados normalizados	4,294	3,770	2,913	3,733

Fonte: autoria própria

Seguindo as métricas de gêneros e artistas, conseguimos ver que em suas médias o valor que utiliza *k-means* com $k = 5$ se mostra superior em relação aos outros agrupamentos, seguindo o valor sugerido pelo método do cotovelo, assim como na análise do método com PCA.

Como o coeficiente de silhueta não está presente aqui, conseguimos encontrar o melhor grupo levando em conta as métricas ano, gênero e artista individualmente.

Em uma análise mais aprofundada do método t-SNE com $k = 5$, os dados indicam que as contagens médias de anos variam conforme o agrupamento. A maior entropia dos anos é evidenciada no grupo 4, com um valor de 5,137, enquanto os outros agrupamentos apresentam valores mais baixos: grupo 2 (4,940), grupo 3 (4,950), grupo 1 (4,868) e grupo 5 (4,733).

Quanto às métricas de gênero e artista, os dados também revelam variações significativas. No que diz respeito ao gênero, o agrupamento 5 apresenta a maior entropia, com um valor de 2,122, superando os demais agrupamentos: grupo 1 (1,795), grupo 2 (1,837), grupo 3 (1,994) e grupo 4 (2,122).

Analogamente, na métrica de artista, o agrupamento 5 também prevalece com a maior média, alcançando 4,378. Os outros agrupamentos apresentam valores de entropia um pouco menores: grupo 1 (4,274), grupo 2 (3,681), grupo 3 (4,229) e grupo 4 (4,130).

Levando em conta o grupo com maior entropia para métrica de ano, conseguimos exemplificar as imagens presentes no grupo um, como vemos na Figura 10.

Figura 10: Exemplo do quarto grupo gerado a partir do método utilizando t-SNE com *k-means* e $k = 5$

Fonte: autoria própria

Levando em conta o grupo com maior entropia considerando as métricas de gênero e artistas, conseguimos exemplificar as imagens presentes no grupo quatro, como vemos na Figura 11.

Figura 11: Exemplo do quinto grupo gerado a partir do método utilizando t-SNE com *k-means* e $k = 5$

Fonte: autoria própria

4.2. ANÁLISE COMPARATIVA

Com base nos resultados apresentados, a análise comparativa entre os métodos de agrupamento utilizados, *k-means* e agrupamento hierárquico, em conjunto com PCA e t-SNE, revela uma variedade de informações sobre a eficácia de cada método.

Para os dois métodos de redução de dimensionalidade - PCA e t-SNE - o método de agrupamento *k-means* com $k = 5$ consistentemente apresentou desempenho superior em todas as métricas de avaliação usadas (ano, gênero, artista e coeficiente de silhueta). O método do cotovelo reiterou este resultado, indicando cinco como o número ideal de agrupamentos para ambos os casos. Assim, o *k-means* com $k = 5$ foi considerado como o método mais eficiente para a categorização das pinturas impressionistas em ambos os casos.

Com base na análise apresentada, é possível destacar que ambos os métodos, PCA e t-SNE, demonstraram eficácia no agrupamento dos dados. No entanto, com base nos dados apresentados, o PCA obteve valores, em relação às métricas de anos, artistas e gêneros com dados normalizados, maiores comparados ao t-SNE.

Para o coeficiente de silhueta, que mede a coesão dos grupos, o PCA apresentou valores mais elevados, o que sugere que ele foi mais eficaz na criação de segmentos mais coesos. Além disso, o PCA com o algoritmo *k-means* e $k = 5$ superou os outros métodos nos valores de entropia dos anos, gêneros e artistas, tanto para os dados normalizados quanto não normalizados.

Em relação ao t-SNE, embora tenha tido resultados favoráveis em várias métricas e tenha se destacado especialmente no agrupamento por gêneros e artistas, a falta do coeficiente de silhueta torna mais difícil a avaliação objetiva da coesão dos grupos criados por esse método. Além disso, apesar de apresentar bons resultados em relação aos gêneros com dados não normalizados, ele foi ligeiramente superado pelo PCA nas métricas gerais de anos e artistas.

Portanto, é possível concluir que, para este grupo de dados específico e as métricas selecionadas para avaliação, o PCA parece oferecer um desempenho geral superior ao do t-SNE. No entanto, é importante lembrar que a escolha do melhor método depende fortemente do conjunto de dados específico e do objetivo da análise. Portanto, embora o PCA tenha demonstrado um desempenho superior neste caso, o t-SNE pode ser a melhor escolha em outras situações.

Ao analisar o agrupamento hierárquico, observou-se que, em comparação com o método *k-means*, a coesão dos grupos era geralmente mais baixa. Isso foi evidente tanto nos coeficientes de silhueta, como nos valores de entropia dos anos, gêneros e artistas. No entanto, o agrupamento hierárquico oferece uma visualização mais intuitiva da hierarquia de agrupamentos, o que pode servir para entender as relações entre diferentes grupos de obras de arte.

5. CONSIDERAÇÕES FINAIS

Através da análise em nosso estudo, foi possível explorar a utilidade e a aplicação de métodos de redução de dimensionalidade e técnicas de agrupamento para avaliação de pinturas impressionistas. Nosso foco foi direcionado aos métodos de PCA e t-SNE, ambos utilizados em conjunto com o método de agrupamento *k-means* e agrupamento hierárquico.

Conforme as análises, o número ideal de agrupamentos para o conjunto de dados em questão foi identificado como sendo cinco, o que foi consistentemente evidenciado pelo método do cotovelo e confirmado por diversas métricas de avaliação, incluindo ano, gênero e artista.

Para o método PCA, a abordagem de agrupamento *k-means* com $k = 5$ apresentou os melhores resultados, como foi evidenciado pelas maiores médias nas métricas de avaliação e pelo coeficiente de silhueta, que indicou uma maior coesão para o primeiro grupo.

A análise do método t-SNE também revelou que o agrupamento *k-means* com $k = 5$ produziu os maiores valores para as métricas de avaliação. Entretanto, sem a capacidade de avaliar o coeficiente de silhueta devido à sua natureza não determinística, aprofundamos a análise nos resultados dos grupos, onde cada um apresentou variações significativas na entropia de cada métrica.

No entanto, é importante ressaltar que a eficácia de qualquer método de agrupamento pode ser altamente dependente da natureza e da complexidade dos dados. Portanto, embora nossas análises tenham produzido resultados claros e consistentes para este conjunto de dados, deve-se ter cuidado ao generalizar essas conclusões para outros conjuntos de dados.

Finalmente, vale destacar que a compreensão de padrões latentes e a extração de informações relevantes de grandes conjuntos de dados são tarefas complexas, mas fundamentais para a tomada de decisões informadas em muitos campos. Através da nossa análise, demonstramos a utilidade de métodos de redução de dimensionalidade e técnicas de agrupamento para facilitar essa tarefa.

Nossos resultados sublinham a importância de abordagens quantitativas robustas na análise de dados e reiteram a necessidade de uma escolha criteriosa e fundamentada dos métodos a serem aplicados, dependendo das especificidades do conjunto de dados em mãos.

5.1. TRABALHOS FUTUROS

Nossa análise focou na utilização de PCA e t-SNE como métodos de redução de dimensionalidade, associados ao *k-means* para o agrupamento dos dados. No entanto, o campo da aprendizagem de máquina é vasto, e muitos outros métodos podem ser explorados para entender ainda mais a complexidade do nosso conjunto de dados.

Um caminho promissor para pesquisas futuras seria a implementação de outras métricas de avaliação de agrupamentos, além do coeficiente de silhueta. Métricas como o Índice de Rand Ajustado, a Informação Mútua Normalizada e a Homogeneidade, Completude e Medida V poderiam ser consideradas para avaliar a qualidade e a estabilidade dos agrupamentos de uma perspectiva diferente. A implementação dessas métricas alternativas poderia revelar insights complementares e contribuir para uma análise mais robusta e completa.

Além disso, a exploração de outras técnicas de agrupamento, como *k-means*++, Agrupamento Espacial Baseado em Densidade de Aplicativos com Ruído (do inglês, *Density-Based Spatial Clustering of Applications with Noise* - DBSCAN), ou algoritmos de agrupamento baseados em densidade, pode trazer novas perspectivas sobre a estrutura do nosso conjunto de dados. O *k-means*++ é uma versão aprimorada do *k-means* cujo objetivo é superar a dependência dos centróides iniciais, enquanto DBSCAN e outros métodos baseados em densidade podem ser particularmente úteis para identificar agrupamentos de formas complexas e para lidar com ruídos e dados discrepantes.

Além disso, considerando que a nossa base de dados atual foca em certos estilos de arte, seria enriquecedor investigar outros estilos artísticos em pesquisas

futuras. A inclusão de mais estilos e épocas, bem como de obras de diferentes partes do mundo, poderia ampliar a diversidade do nosso conjunto de dados e permitir uma análise mais abrangente e inclusiva.

REFERÊNCIAS

[1] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision. (2015)

[2] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. How transferable are features in deep neural networks? (2014)

[3] Jian Xiao, Jia Wang, Shaozhong Cao, Bilong Li. Application of a Novel and Improved VGG-19 Network in the Detection of Workers Wearing Masks. Journal of Physics: Conference Series. (2020)

[4] Dario Garcia-Gasulla, Ferran Parés, Armand Vilalta, Jonatan Moreno, Eduard Ayguadé, Jesús Labarta, Ulises Cortés, Toyotaro Suzumura. On the Behavior of Convolutional Nets for Feature Extraction. Journal of Artificial Intelligence Research. (2017)

[5] Jolliffe, I. T., & Cadima, J. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. (2016)

[6] Van der Maaten, L., & Hinton, G. Visualizing data using t-SNE. Journal of machine learning research. (2008)

[7] MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. (1967)

[8] Jain, A. K., & Dubes, R. C. Algorithms for clustering data. Prentice-Hall, Inc. (1988)

[9] Annibale Panichella, Bogdan Dit, Rocco Oliveto, Massimiliano Di Penta, Denys Poshynanyk, Andrea De Lucia. How to Effectively Use Topic Models for Software Engineering Tasks? An Approach Based on Genetic Algorithms. Proceedings - International Conference on Software Engineering (2013)

[10] Hestry Humaira & Rasyidah Rasyidah. Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm. Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia (2020)

[11] Shannon, C. E. A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379-423. (1948)

[12] MacKay, D. J. C. Information Theory, Inference, and Learning Algorithms. Cambridge University Press. (2003)

[13] Cover, T. M., & Thomas, J. A. Elements of information theory (Vol. 6). John Wiley & Sons. (2006)

[14] Eren Gultepe, Thomas E. Conturo, Masoud Makrehchi. Predicting and Grouping Digitized Paintings by Style using Unsupervised Feature Learning. Journal of Cultural Heritage. (2018)

[15] Giovanna Castellano, Gennaro Vessio. A Deep Learning Approach to Clustering Visual Arts. Computer Vision and Cultural Heritage Preservation. (2022)

[16] Giovanna Castellano, Gennaro Vessio. Deep Convolutional Embedding for Digitized Painting Clustering. 25th International Conference on Pattern Recognition. (2020)

[17] Ville Satopa, Jeannie Albrecht, David Irwin, Barath Raghavan. Finding a “Kneedle” in a Haystack: Detecting Knee Points in System Behavior. Distributed Computing Systems Workshops (ICDCSW). (2011)