Classificação e cariotipagem de imagens de Cromossomos Humanos com Redes Neurais Convolucionais e Aprendizado de Conjunto - Nível: Doutorado.

Francisco das C. Imperes Filho[1], Vinicius P. Machado[1], Arlino Magalhães[1] and Rodrigo de M. S. Veras[1]

[1]*Núcleo de Computação de Alto Desempenho – UFPI – Teresina – PI – Brazil*
{fcoimperes,vinicius, arlino, rveras}@ufpi.edu.br
1

Abstract. *Chromosome classification constitutes an essential undertaking in the identification of genetic anomalies, a task conventionally executed by geneticists via manual karyotype analysis. This study puts forth an approach to leverage pre-trained Convolutional Neural Networks for the development of a model. This model utilizes images featuring overlapping chromosomes to simulate clinical scenarios and is designed to execute the classification of chromosomes into their respective karyotype classes derived from microscopic imagery. The ultimate objective is the development of a system capable of the automated identification, classification, and assembly of karyotypes from complex images, concurrently investigating the potential of Ensemble Learning and interpretability tools, such as Saliency Maps.*

Resumo. *A classificação de cromossomos é uma tarefa essencial na identificação de anomalias genéticas, tradicionalmente realizada por geneticistas mediante análise manual de cariótipos. Este trabalho propõe uma abordagem para explorar Redes Neurais Convolucionais pré-treinadas para desenvolver um modelo, utilizando imagens com cromossomos sobrepostos para simular cenários clínicos, capaz de realizar o pareamento de cromossomos em suas respectivas classes de cariótipo a partir de imagens microscópicas. O objetivo final é desenvolver um sistema que realize a identificação, classificação e montagem automática de cariótipos a partir de imagens complexas, investigando também o potencial do Ensemble Learning e de ferramentas de interpretabilidade, como Saliency Maps.*

1. Informações Gerais

1. Nome do estudante: Francisco das Chagas Imperes Filho - fcoimperes@ufpi.edu.br.
2. Orientador: Vinicius P. Machado – vinicius@ufpi.edu.br
3. Coorientador: Prof. Dr. Rodrigo de Moraes Soares Veras - rveras@ufpi.edu.br.
4. Universidade/ Programa: Universidade Federal do Piauí (UFPI) / Programa de Pós-Graduação em Ciência da Computação (DCCMAPI).
5. Mês de ingresso / Previsão de defesa: AGO/2023 - JUL/2027.
6. Etapas concluídas e datas de referência:
   a) Disciplinas: 2024.2 - Qualificação: 2025.1
   b) Proposta de Tese: 2026.1

2. Introdução e Motivação

A cariotipagem[1] é uma prática essencial na identificação de doenças genéticas. No entanto, o processo manual de classificação de cromossomos é lento, exige especialistas treinados e está sujeito a erros [Anh et al. 2022]. Esses desafios evidenciam a necessidade de soluções automatizadas que melhorem a precisão, agilidade e padronização do diagnóstico. Nesse contexto, o avanço da Inteligência Artificial (IA), especialmente por meio do uso de Redes Neurais Convolucionais (CNN), tem possibilitado novas abordagens na área biomédica, permitindo a difusão de métodos para reduzir a carga de trabalho humano e a padronização de diagnósticos [Saranya and Lakshmi 2023].

Diante deste cenário, este trabalho propõe desenvolver um modelo de CNNs para classificação de cromossomos humanos, visando superar as limitações atuais. A motivação é criar uma solução que seja eficaz, interpretável e aplicável clinicamente, especialmente para cromossomos sobrepostos e para automatizar a montagem do cariótipo. Adicionalmente, por lidar com a classificação e a recuperação de grandes volumes de imagens biomédicas, a proposta se insere na área de Banco de Dados, visto que a organização e o gerenciamento desses dados, associados aos metadados clínicos, exigem soluções que integrem mineração de dados e aprendizado de máquina [Magalhães et al. 2023].

A inovação reside em integrar a IA aplicada à biomedicina com o gerenciamento inteligente de dados. O sistema otimizará o acesso e a gestão de grandes volumes de imagens biomédicas e metadados clínicos, facilitando a recuperação e análise em larga escala, e posicionando-se na interseção entre IA e Banco de Dados.

3. Trabalhos Relacionados

Nesta seção, apresentamos trabalhos relevantes na área de classificação de cromossomos.

O trabalho de [Lin et al. 2022] teve como objetivo construir, reproduzir, avaliar e comparar modelos e algoritmos de classificação de cromossomos existentes em um conjunto de dados de imagens de cromossomos. Segundo os autores, o conjunto permite uma avaliação comparativa e a construção de linhas de base de classificação cromossômica adequada para diferentes cenários. O conjunto de imagens consiste em 126.453 instâncias de cromossomos corados em Banda-G, preservando a privacidade de 2.763 cariótipos de 408 indivíduos. Os pesquisadores ressaltam que a melhor linha de base com 0,99% de precisão, *recall* e *F1-score* e com 99,33% de acurácia, relata desempenho de classificação de última geração.

O estudo de [Xia et al. 2023], intitulado *KaryoNet*, introduziu uma metodologia para o reconhecimento de cromossomos, focando na identificação de interações de longo alcance entre cromossomos em células metafásicas[2] para aprimorar a classificação de tipos e polaridades cromossômicas em cariótipos normais e anormalidades numéricas. Empregando uma abordagem que incluiu técnicas de extração de características via CNN, *Masked Feature Interaction Module* (MFIM), *Deep Assignment Module* (DAM), *ResNet50*, *Transformer encoder*, *Bi-RNN blocks* e *Gated Recurrent Unit* (GRU), o

[1]Método que analisa os cromossomos de uma célula para identificar possíveis alterações genéticas, seja em número, tamanho, forma ou estrutura.

[2]Células metafásicas são um estágio específico no ciclo celular, ocorrendo após a prófase.

KaryoNet alcançou acurácia de 98,41% para cromossomos *R-band* e 99,58% para cromossomos *G-band* em avaliações clínicas, superando o *software* comercial *Ikaros* com uma acurácia registrada de 90,17%.

Contudo, os trabalhos mencionados focam em cromossomos isolados ou em suas interações, não abordando imagens clínicas complexas com sobreposição ou aglomeração. Gerenciar esses cenários para a montagem automática de cariótipos a partir de imagens complexas é o foco principal da nossa proposta. Adicionalmente, buscaremos aprimorar a interpretabilidade dos resultados para validação clínica, um ponto deficiente nas abordagens atuais.

4. Problema

A classificação manual de cromossomos ainda é uma prática predominante em muitos laboratórios de genética, demandando tempo e expertise do geneticista e sendo suscetível a inconsistências. Apesar dos avanços tecnológicos, a implementação de sistemas automatizados ainda enfrenta obstáculos importantes. Os principais desafios estão relacionados à variabilidade morfológica dos cromossomos, à similaridade entre classes vizinhas e à complexidade em capturar padrões discriminativos em imagens com baixa resolução ou ruído.

Ademais, a maioria das abordagens existentes não explora de forma integrada aspectos como consistência estatística, interpretabilidade dos modelos e avaliação em grandes volumes de dados. Embora existam soluções promissoras, como o *KaryoNet* [Xia et al. 2023] e outras redes convolucionais, muitas dessas soluções são limitadas pela indisponibilidade dos dados, ausência de validação cruzada ou pela falta de métodos de avaliação complementares (como o índice *Kappa* e desvio padrão), que são cruciais para garantir a confiabilidade dos sistemas em contextos clínicos reais.

Para superar os desafios identificados, o problema central deste trabalho consiste em desenvolver um modelo automatizado para classificar cromossomos em suas respectivas classes de cariótipo, diretamente de imagens microscópicas, eliminando a necessidade de segmentação prévia. Embora a segmentação seja uma etapa comum em muitos modelos de análise de imagem, essa abordagem visa simplificar o processo, reduzir erros e agilizar o fluxo de trabalho clínico. O modelo deve ser interpretável e aplicável em laboratórios, superando a complexidade de imagens, variabilidade de amostras e dificuldade de interpretação das abordagens atuais.

5. Solução Proposta

Com o intuito de consolidar a aplicabilidade clínica do problema apontado, esta proposta fundamenta-se em três etapas cruciais.

5.1. Modelo baseado em CNNs pré-treinadas.

A primeira etapa propõe uma abordagem baseada em CNNs pré-treinadas, combinada com técnicas de *Deep Fine Tuning* (DFT), *data augmentation*, *cross-validation k-fold* para testar algumas arquiteturas pré-treinadas, a fim de identificar a mais adequada para a tarefa de classificação cromossômica.

Além da comparação entre os otimizadores[3] *Adaptive Moment Estimation* (ADAM) e *Stochastic Gradient Descent* (SGD), também será usada a técnica de explicabilidade de mapas de saliência *(Saliency Maps)* para analisar visualmente o processo de decisão da rede. Nosso objetivo não é desvendar completamente o funcionamento interno do *deep learning*, mas sim oferecer explicações úteis ao usuário final, aumentando a confiança no uso clínico do modelo.

Para elevar a precisão e a consistência dos resultados, empregaremos o aprendizado de conjunto *(ensemble learning)*. Combinaremos múltiplas CNNs, selecionadas por seu desempenho individual e diversidade, usando uma média ponderada. Isso permitirá que o modelo generalize melhor para diferentes tipos de imagens de cromossomos.

Finalmente, a performance do modelo será avaliada não só por métricas como acurácia, precisão, *recall* e *F1-Score*, mas também pelo coeficiente *Kappa* (essencial em cenários com classes desequilibradas) e pelo desvio padrão. Essas métricas garantirão que o modelo seja confiável e estável em diversas condições de dados.

5.2. Base de Imagens com Cromossomos Sobrepostos ou Aglomerados

Esta etapa consiste na construção e utilização de um conjunto de dados contendo imagens que simulam condições clínicas reais mais complexas, nas quais os cromossomos aparecem parcialmente sobrepostos ou agrupados. Diferentemente das amostras isoladas, imagens sobrepostas ou agrupadas refletem o desafio típico encontrado em ambientes laboratoriais, caracterizados por ruído, baixa padronização e variações morfológicas pronunciadas.

A motivação para esta abordagem baseia-se no reconhecimento de que modelos treinados exclusivamente em dados idealizados tendem a sofrer perda significativa de desempenho quando aplicados a imagens reais. Ao introduzir essas variações, espera-se aumentar a consistência do sistema, aprimorar sua capacidade de generalização e torná-lo apto a operar com confiabilidade em contextos clínicos, onde a qualidade e a padronização das amostras podem ser limitadas.

Para tanto, serão realizadas atividades como a coleta, o pré-processamento e/ou a simulação de imagens contendo cromossomos sobrepostos, a adaptação do modelo de treinamento desenvolvido na primeira etapa da proposta para lidar com estas características e a avaliação do impacto dessas imagens na performance do modelo, incluindo métricas e índices que avaliem a consistência estatística.

5.3. Montagem Automática de Cariótipos com Classificação Integrada

Na terceira etapa, propõe-se o desenvolvimento de um sistema automatizado que englobe todo o fluxo de análise cromossômica a partir de imagens complexas, capazes de conter múltiplos cromossomos. O sistema será responsável por identificar, classificar e agrupar os cromossomos para formar automaticamente os 23 pares cromossômicos característicos do cariótipo humano.

A automação desta etapa responde a uma necessidade crítica dos laboratórios clínicos, onde a montagem manual de cariótipos é um processo lento, trabalhoso e

[3]Algoritmos que têm a finalidade de minimizar a função de perda (erro) da rede, ajustando os pesos e vieses (parâmetros) da rede neural durante o processo de treinamento.

sujeito a falhas. Ao integrar CNNs e algoritmos de agrupamento para a montagem dos pares cromossômicos, a proposta visa proporcionar uma ferramenta capaz de acelerar significativamente a análise, reduzindo custos e aumentando a precisão diagnóstica.

Outrossim, o sistema incorporará mecanismos de explicabilidade visual, por meio da geração de mapas de calor, possibilitando interpretações claras e auditáveis das decisões do modelo. Tal funcionalidade é fundamental para garantir a confiança dos profissionais da saúde na aplicação clínica do sistema, facilitando sua adoção em ambientes laboratoriais e hospitalares.

5.4. Impactos e Benefícios Esperados

A execução das etapas anteriores representa um avanço na implementação de sistemas automatizados de análise cromossômica. Ao lidar com imagens complexas e automatizar a montagem de cariótipos, o modelo baseado em CNNs:

- Aumentará a confiabilidade da classificação cromossômica;
- Reduzirá o tempo e custo necessários para a análise genética;
- Fornecerá ferramentas para auxiliar na interpretabilidade e na auditabilidade das decisões do modelo, promovendo uma maior confiança e aceitação clínica; reforçando que, embora as CNNs sejam 'caixas pretas', as ferramentas de interpretabilidade visual são um passo fundamental para a adoção clínica.
- Possibilitará a integração com sistemas hospitalares, configurando-se como um recurso para triagem e diagnóstico genético.

Faremos uma análise do custo computacional (treinamento e inferência) para otimizar a eficiência do sistema. Isso garante sua aplicabilidade em ambientes com recursos variados, atendendo às necessidades práticas e impulsionando a análise, classificação e montagem automatizada de cariótipos humanos.

6. Contribuições da Proposta de Tese

Com o objetivo de automatizar e aprimorar a análise de cariótipos, este trabalho visa às seguintes contribuições:

1. Desenvolver um sistema que identifica, classifica e organiza cromossomos em cariótipos a partir de imagens sobrepostas.
2. Melhorar a classificação de cromossomos usando CNNs e Aprendizado de Conjunto.
3. Tornar os modelos de IA mais transparentes para geneticistas através de ferramentas como Mapas de Saliência, aumentando a confiança clínica.
4. Garantir a consistência e confiabilidade dos resultados do modelo usando múltiplas métricas, incluindo *Kappa* e desvio padrão.
5. Contribuir com soluções para organizar, classificar e acessar grandes volumes de imagens biomédicas e dados associados.
6. Comparar o desempenho do modelo proposto com a classificação manual, quantificando melhorias em precisão, tempo e redução de inconsistências.

7. Avaliação preliminar dos resultados

A seguir, apresentamos um resumo dos resultados obtidos na primeira etapa do modelo. Os experimentos iniciais foram realizados sobre o conjunto de imagens *ChromosomeNet* proposto por [Lin et al. 2022]. Detalhes sobre o *dataset* utilizado podem ser encontrados na Seção 3). Embora as imagens iniciais deste *dataset* sejam de cromossomos isolados, o projeto visa, em etapas seguintes, trabalhar com imagens simulando cenários clínicos reais com cromossomos sobrepostos ou aglomerados, para assegurar a aplicabilidade do modelo em condições mais complexas, que refletem as amostras obtidas em ambientes laboratoriais.

Descartamos 1.234 imagens do conjunto *ChromosomeNet* porque não atendiam aos critérios de isolamento de cromossomos necessários para a primeira etapa do trabalho. Muitas dessas imagens continham múltiplos cromossomos por imagem ou apenas um ponto sem relevância para a análise, o que as tornava inadequadas para o nosso estudo.

A Tabela 1 expressa que as arquiteturas *DenseNet169* e *Xception* obtiveram os melhores resultados. A primeira teve desempenho levemente superior atingindo acurácia, *recall* e *F1-Score* de 98,77%, e precisão de 98,78%. O coeficiente *Kappa* foi de 0,99, indicando uma concordância quase perfeita, e o desvio padrão das métricas foi de 0,002 para as duas arquiteturas, demonstrando alta estabilidade nos resultados. Para uma análise detalhada de todos os resultados e discussões, o trabalho completo está disponível em [Imperes et al. 2024].

Tabela 1. Resultados dos experimentos utilizando a estratégia DFT com otimizador ADAM.

Arquitetura	A (%)	P (%)	R (%)	F1-Score (%)	K
DenseNet121	98,74 ± 0,002	98,75 ± 0,002	98,74 ± 0,002	98,74 ± 0,002	0,99 ± 0,002
DenseNet169	**98,77 ± 0,002**	**98,78 ± 0,002**	**98,77 ± 0,002**	**98,77 ± 0,002**	**0,99 ± 0,002**
DenseNet201	98,42 ± 0,005	98,46 ± 0,005	98,42 ± 0,005	98,42 ± 0,005	0,98 ± 0,005
EfficientNetV2B3	98,74 ± 0,001	98,75 ± 0,001	98,74 ± 0,001	98,74 ± 0,001	0,99 ± 0,001
InceptionV3	98,75 ± 0,002	98,76 ± 0,002	98,75 ± 0,002	98,75 ± 0,002	0,99 ± 0,002
Resnet50	98,26 ± 0,005	98,28 ± 0,005	98,26 ± 0,005	98,26 ± 0,005	0,98 ± 0,006
VGG-16	98,09 ± 0,005	98,12 ± 0,005	98,09 ± 0,005	98,09 ± 0,005	0,98 ± 0,006
VGG-19	98,09 ± 0,003	98,11 ± 0,003	98,09 ± 0,003	98,09 ± 0,003	0,98 ± 0,003
Xception	**98,76 ± 0,002**	**98,77 ± 0,002**	**98,76 ± 0,002**	**98,76 ± 0,002**	**0,99 ± 0,002**

8. Estado Atual do Trabalho

Para fornecer um panorama do andamento do trabalho, as atividades que estão sendo realizadas no projeto de pesquisa são apresentadas a seguir:

1. Revisão bibliográfica sobre abordagens automatizadas de classificação de imagens de cromossomos (em andamento).
2. Especificação das técnicas e métricas para avaliação da proposta de classificação (concluída).
3. Desenvolvimento de um modelo para classificação de imagens de cromossomos utilizando CNNs pré-treinadas (concluída).
4. Produção da tese e seleção de novos conjuntos de imagens simulando situações clínicas laboratoriais (iniciadas).
5. Desenvolvido um sistema para identificação, classificação e montagem de cariótipos humanos (não iniciada).

9. Conclusão

Os experimentos de classificação de imagens de cromossomos revelaram que a performance inicial do modelo CNN é significativamente influenciada pela escolha da arquitetura da rede e do otimizador utilizados. A estratégia DFT mostrou-se promissora, independentemente do otimizador usado, com todas as arquiteturas alcançando métricas acima de 90%.

A escolha do otimizador teve um impacto significativo na performance dos modelos. Especificamente, o otimizador *ADAM* com a estratégia DFT demonstrou promover maior precisão e estabilidade nos resultados. Isso sugere que a utilização deste otimizador pode ser eficaz para melhorar a eficiência do modelo em tarefas de classificação complexas.

Como limitações, destaca-se a necessidade de testes e validação em outras bases de imagens, com cromossomos sobrepostos, além da integração com sistemas de montagem automática de cariótipos, proposta como etapas futuras deste projeto.

Referências

[Anh et al. 2022] Anh, L., Thanh, V., Son, N., Phuong, D. K., Anh, L., Ram, T., Minh, B., Tung, H., Thinh, N., Ha, V., and Ha, M. (2022). Efficient type and polarity classification of chromosome images using cnns: a primary evaluation on multiple datasets. In *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE)*, pages 400–405.

[Imperes et al. 2024] Imperes, F. D. C. F., Machado, V. P., do Monte, S. J. H., dos Santos, A. R., de Sousa, L. P., da Silva, A. S., Pereira, E. M., and Veras, R. M. S. (2024). Explorando arquiteturas de redes neurais profundas na classificação de imagens de cariótipos humanos. *Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2024)*.

[Lin et al. 2022] Lin, C., Chen, H., Huang, J., Peng, J., Guo, L., Yang, Z., Du, J., Li, S., Yin, A., and Zhao, G. (2022). Chromosomenet: A massive dataset enabling benchmarking and building basedlines of clinical chromosome classification. *Computational Biology and Chemistry*, 100:107731.

[Magalhães et al. 2023] Magalhães, A., Monteiro, J., and Ângelo Brayner (2023). Main memory database instant recovery. In *Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados (SBBD)*, pages 255–269, Porto Alegre, RS, Brasil. SBC.

[Saranya and Lakshmi 2023] Saranya, S. and Lakshmi, S. (2023). Classification of chromosomes to diagnose chromosomal abnormalities using cnn. In *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, pages 1–5.

[Xia et al. 2023] Xia, C., Wang, J., Qin, Y., Wen, J., Liu, Z., Song, N., Wu, L., Chen, B., Gu, Y., and Yang, J. (2023). Karyonet: Chromosome recognition with end-to-end combinatorial optimization network. *IEEE Transactions on Medical Imaging*, 42(10):2899–2911.