

Mateus Taulois Braga

# **Projeto Orientado em Computação**

Belo Horizonte, Minas Gerais

2022

Mateus Taulois Braga

## **Projeto Orientado em Computação**

Detecção de Proteínas Monoclonais em  
Exame de Imunofixação

Universidade Federal de Minas Gerais  
Instituto de Ciências Exatas  
Departamento de Ciência da Computação

Orientador: Dr. Flavio Vinicius Diniz de Figueiredo

Belo Horizonte, Minas Gerais  
2022

# Sumário

1	ABSTRACT . . . . .	3
2	INTRODUÇÃO . . . . .	4
2.1	Objetivos Gerais . . . . .	5
2.2	Objetivos Específicos . . . . .	5
3	REFERENCIAL TEÓRICO . . . . .	7
4	METODOLOGIA . . . . .	8
5	RESULTADOS OBTIDOS . . . . .	13
6	CONCLUSÃO . . . . .	16
	REFERÊNCIAS . . . . .	17

# 1 Abstract

Este trabalho aborda a automatização do processo diagnóstico e seus benefícios para a obtenção de resultados confiáveis em tempo mínimo. O estudo foi realizado no laboratório Hermes Pardini, onde cerca de 20.000 exames mensais são filtrados de forma autônoma. A pesquisa destacou que o tempo de formação de um profissional apto a realizar a análise processual e fornecer resultados em cinco a dez minutos demanda aproximadamente um ano. Em contrapartida, após o desenvolvimento do programa de automatização, o tempo de laudo é reduzido para segundos. O principal objetivo desse trabalho é propor uma solução efetiva para agilizar o processo diagnóstico e entregar resultados confiáveis em tempo mínimo.

## 2 Introdução

A automatização no processo diagnóstico está trazendo grandes benefícios ao laboratório Hermes Pardini, permitindo que cerca de 20.000 exames mensais sejam filtrados de forma autônoma. Essa automação garante uma maior rapidez na obtenção dos resultados. Essa melhoria significativa se deve a uma série de razões, sendo uma delas o tempo necessário para formar um profissional capaz de realizar a análise processual e fornecer o resultado em um intervalo de cinco a dez minutos, que leva em média um ano.

Em contrapartida, uma vez desenvolvido, o programa de automatização consegue emitir o laudo em questão de segundos. Essa diferença de tempo é notável e evidencia um dos principais objetivos desse trabalho: propor uma solução efetiva que entregue resultados confiáveis em um tempo mínimo.

A redução significativa do tempo necessário para emitir os laudos dos exames traz inúmeras vantagens. Em primeiro lugar, os pacientes podem receber seus resultados mais rapidamente, permitindo que sejam tomadas ações imediatas caso necessário. Além disso, o laboratório consegue aumentar sua capacidade de atendimento, processando um maior número de exames em um menor período de tempo.

A confiabilidade dos resultados também é um fator crucial nesse contexto. A automatização do processo diagnóstico garante uma padronização e consistência nas análises, minimizando erros humanos e aumentando a precisão dos resultados. Com a utilização de algoritmos e inteligência artificial, é possível realizar análises mais complexas e identificar padrões que podem ser essenciais para um diagnóstico preciso.

Portanto, o desenvolvimento desse trabalho representa um avanço significativo na área da saúde, proporcionando uma solução efetiva para agilizar o processo diagnóstico e entregar resultados confiáveis em tempo mínimo. A automatização no laboratório Hermes Pardini demonstra o potencial da tecnologia em melhorar a eficiência e a qualidade dos serviços prestados, contribuindo para o bem-estar dos pacientes e aprimorando a medicina diagnóstica como um todo.

No *pipeline* de diagnóstico, o exame de imunofixação é o segundo, após a avaliação de eletroforese, o qual é apresentado na lâmina como coluna "ELP"e, portanto, não deve ser analisada como proteína.

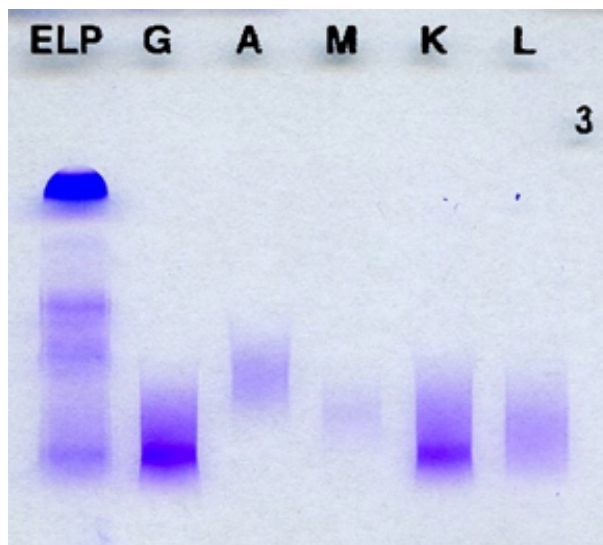


Figura 1 – Exame de Imunofixação, positivo em G e K

## 2.1 Objetivos Gerais

Com a automatização no processo diagnóstico, cerca de 20.000 exames ao mês do laboratório Hermes Pardini serão filtrados de forma autônoma, o que garante maior rapidez ao resultado. Isso se deve, dentre outras, à seguinte razão: o tempo de formação de um profissional a ser apto na análise processual em questão e fornecimento de resultado em cinco a dez minutos gira em torno de um ano. Em contrapartida, depois de desenvolvido, o programa possui um tempo de laudo de segundos. Assim sendo, percebe-se que um dos intuitos com o desenvolvimento deste trabalho consiste em propor uma solução efetiva e que entregue resultados confiáveis em tempo mínimo.

Além disso, é importante ressaltar a explicabilidade do modelo utilizado, ou seja, como é realizada a detecção no exame. É notável que a aceitação do sistema será muito melhor quando os profissionais de saúde compreenderem, em linhas gerais, o funcionamento do sistema. Para tal, realizamos reuniões e alinhamentos periódicos com o usuário final, a fim de trocar informações, que se tornaram cruciais para realizar esta etapa de evidenciar o procedimento por trás do modelo de aprendizado de máquina.

## 2.2 Objetivos Específicos

O objetivo deste projeto é atuar na segunda etapa da linhagem diagnóstica de patologias relacionadas à má absorção proteica, após o exame de *ELP*. Dessa forma, otimiza-se a sequência de análise dos resultados após a triagem de resultados que possuem diagnóstico relevante a ser incluído na sequência do pipeline. Assim sendo, poupa-se a avaliação desnecessária de cerca de 14.000 exames ao mês, os quais serão eliminados de forma segura automaticamente por meio de inteligência artificial.

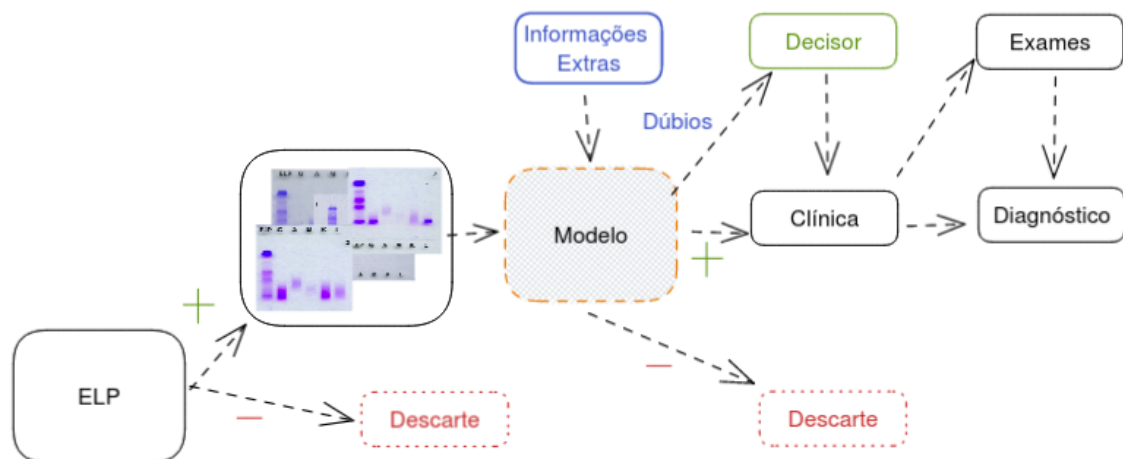


Figura 2 – Fluxograma do diagnóstico

O objetivo do projeto orientado em computação I (POC1) consistiu, portanto, em entender o impacto que variáveis como idade e sexo geram no resultado da testagem para cada tipo específico de proteína, bem como a estruturação de um modelo inicial de predição dos exames.

Como objetivo final, a ideia foi entregar um modelo com, pelo menos 91% de acurácia, que é o acerto dos médicos do laboratório, e focar no ponto de explicabilidade do modelo. Não obstante, foi realizada também a homologação no sistema da Hermes Pardini, bem como a comunicação do laudo do modelo para um laudo disponibilizado para o paciente, de forma online.

### 3 Referencial Teórico

Há pouca literatura disponível quanto ao assunto do uso de modelos de visão computacional como ferramenta de auxílio para interpretação de exames de imunofixação. Apesar de ser um exame amplamente utilizado no meio biológico, a maioria dos esforços científicos são voltados para exames diagnósticos mais urgentes, como câncer e acidente vascular encefálico. Todavia, o estudo [Wei et al. 2021] traz abordagens pertinentes que merecem reflexão aprofundada, tal como a separação por proteínas e um modelo de visão computacional. Entretanto, neste projeto não serão adotados a mesma base de dados e rede, mas terá como apoio o estudo de vetor de características, que irá atuar como decisor.

O trabalho de [Zhao et al. 2018] mostra o uso da rede VGG-16, modelo que será utilizado na predição dos exames. Ela se mostra eficaz em detecção de imagens e possui arquitetura relativamente simples em sua construção.

Ademais, [Ebrahimzadeh et al. 2018] utiliza feature vectors para a classificação de exames médicos, percurso adotado neste projeto.



## 4 Metodologia

Inicialmente, o problema de detecção da presença de uma proteína monoclonal pode ser interpretado como duas classes, e não multi-classe como é apresentado no diagnóstico. O laudo final pode ser qualquer permutação entre todas as proteínas (GK, G, K, AGK, etc.). Porém, como o problema de todo, ou a maioria, dos projetos de ciência de dados e visão computacional, faltariam dados para realizar a predição multi-classe dessa forma. Para tal, será abordada a segmentação das diversas proteínas e classificação binária entre elas, compilando um resultado final com todas as positivas.

O primeiro passo foi a realização de um estudo de filtros, que se demonstrou ineficaz para salientar as marcações positivas nos testes. Evidenciado na figura 3, nota-se que ele realiza a amplificação de proteínas que tem uma facilidade maior em ser detectadas, mas apaga proteínas mais brandas, portanto não sendo utilizado para auxílio na detecção (esquerda e direita, respectivamente).

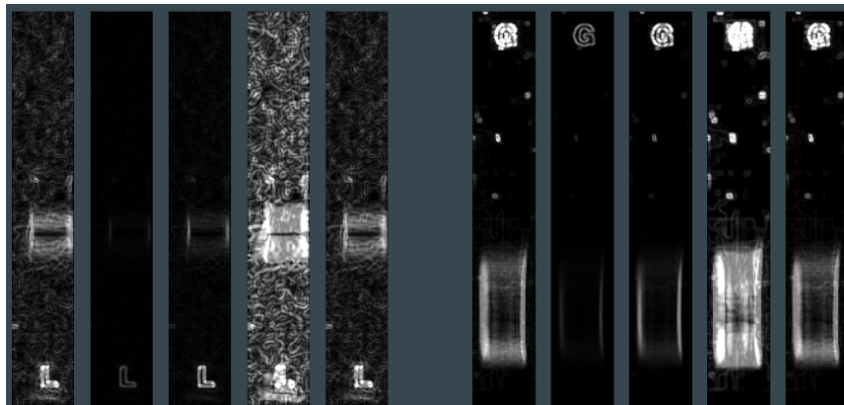


Figura 3 – Filtro de Magnitude Orientation nas proteínas

Foi realizado o treino de um modelo de aprendizado de máquina, com a rede neural VGG16. Todavia, apesar dos resultados se demonstrarem ser suficientemente bons, foi notada uma barreira de aceitação por parte dos médicos e outros profissionais que utilizam o sistema. Enquanto profissionais de área, muitas vezes temos dificuldade de agregar valor ao produto que está sendo desenvolvido, pois o cliente não consegue assimilar o trabalho realizado. Explicar para um pessoa que não tem conhecimento básico sobre computação é difícil, e não garante aceitação do mesmo.

Foi necessária, então, uma mudança de paradigma, para auxiliar a compreensão do usuário final, e melhorar sua aceitação. A partir desse momento, o trabalho focou em realizar uma previsão, embora correta e segura, condizente com as técnicas que são utilizadas pelos profissionais da área médica e biológica. Desta forma, foi realizada a previsão a partir de uma sequência de números, que correspondem a intensidade de cor

Acurácia - F1	Precisão	Recall	F1
88%	87%	86%	87%

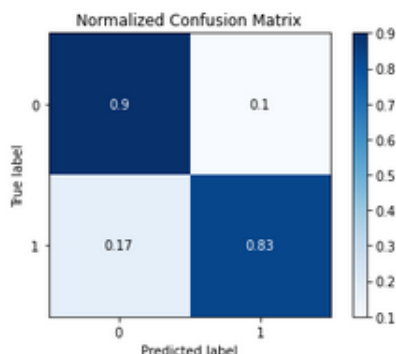


Figura 4 – Resultados com a VGG 16

em cada pixel. Essa sequência foi extraída a partir de uma linha, do topo ao final de cada proteína, como ilustrada na figura seguinte.

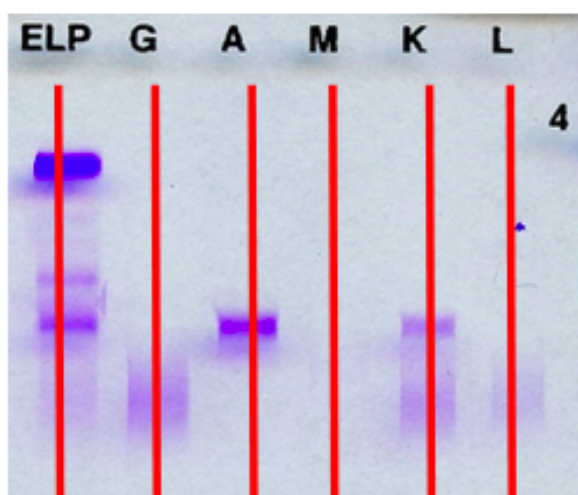


Figura 5 – Exemplo da Extração da Sequência Numérica que é Entrada do Modelo

A partir dessa sequência, foi realizada uma comparação para ver se existe um padrão que define um positivo, portanto foi realizado um plot de comparação das linhas. A linha do ELP sinaliza diversos picos que têm um padrão de falso positivo, uma vez que sua derivada é nitidamente mais acentuada, enquanto a marcação de um verdadeiro positivo é bem mais suave, como nota-se na parte final do plot.

Não obstante, nota-se que uma marcação positiva é localizada em região próxima de outras. Como exemplo, na figura acima, nota-se a presença de três proteínas positivas na região de 160 a 180 pixels. Este padrão se mostrou presente em outras análises, e começou um estudo de análise da localização de correlação entre a localização dos positivos entre proteínas.

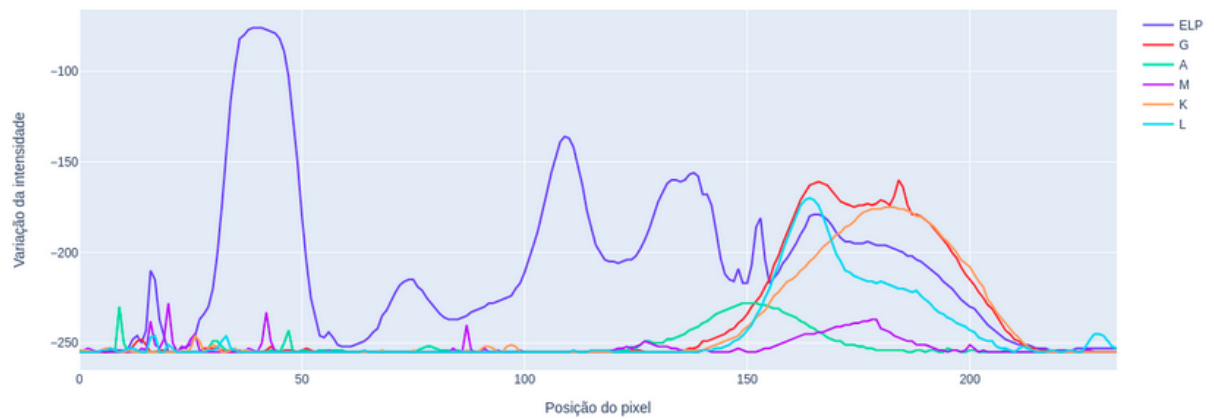


Figura 6 – Plot das Linhas nas Proteínas

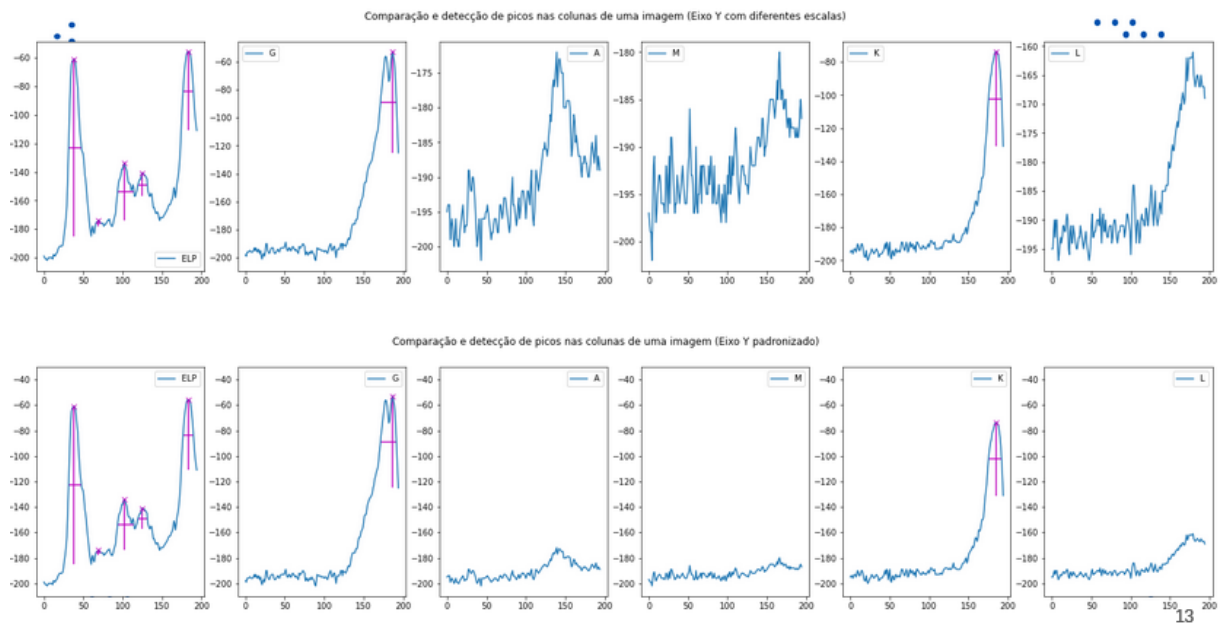


Figura 7 – Plot das Linhas nas Proteínas Separadas, como Extraído e Normalizado

A figura 7, mostra que existe relação entre os picos detectados no ELP e entre outras proteínas positivas. Um pico no ELP marca uma região provável de positivo em outras proteínas. Na segunda linha, com o plot normalizado, fica nítido que existe essa relação com a localização.

Na figura 9, notamos que a marcação não necessariamente implica em uma tradução positiva em outras proteínas, porém mostra que ali é uma região provável, como notamos no exemplo positivo. Com isso, na prática ele é utilizado como um decisor pelo modelo, uma vez que é o primeiro excluidor de um laudo que seria de falso positivo.

Esta mudança de metodologia, de um modelo convencional para um modelo que realiza a predição com base em uma análise específica do problema, resultou em uma melhor comunicação com o cliente, bem como em um resultado com um valor agregado muito

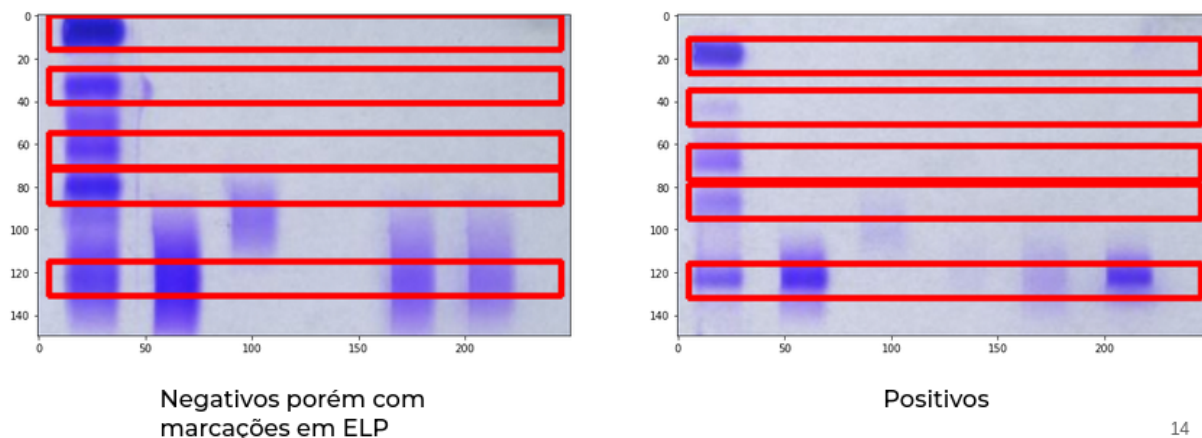


Figura 8 – Exemplo de Localização do ELP e Outras Proteínas

maior, uma vez que se baseia em um modelo que veio do fruto de uma análise específica do problema. A explicabilidade de um modelo é extremamente importante no contexto atual, uma vez que estamos com poder computacional cada vez mais alto e barato, é muito fácil realizar uma predição satisfatória com modelos black box, hospedando uma máquina virtual em algum ambiente cloud, porém o grande diferencial é justamente entender o problema e realizar uma solução específica para a dor do cliente, e, como no caso do modelo atual, utilizar dos dois universos para realizar uma predição ainda melhor.

Não obstante, foi realizada uma detecção de OCR para melhorar a segmentação das proteínas, uma vez que o modelo antigo era com base em detecção de contornos na imagem. Isso resultava em detecções erradas, e gerava segmentações com erros, influenciando negativamente no resultado. Foi utilizada a biblioteca pytesseract para a detecção OCR.

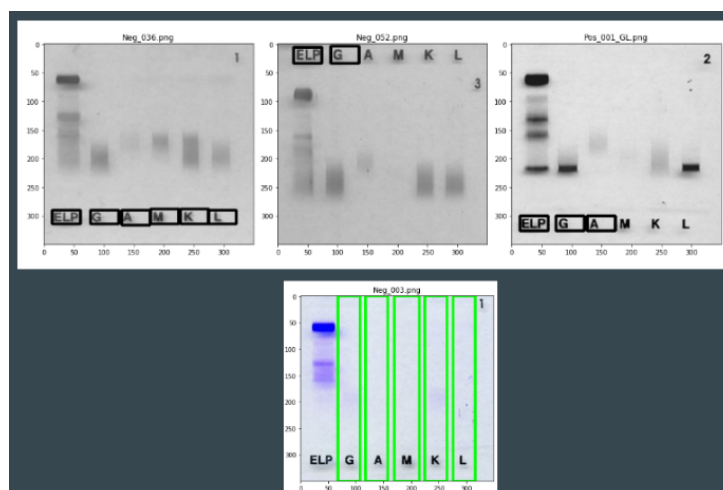


Figura 9 – Exemplo de detecção de OCR para segmentação das proteínas

Com isso, o modelo atual está com uma acurácia ainda maior do que a antiga, como apresentada na matriz de confusão 10.

Não obstante, foi realizado um estudo, proposto inicialmente no primeiro relatório do

Acurácia	F1	Precisão	Recall	True Negative
91%	91%	92%	90%	96%

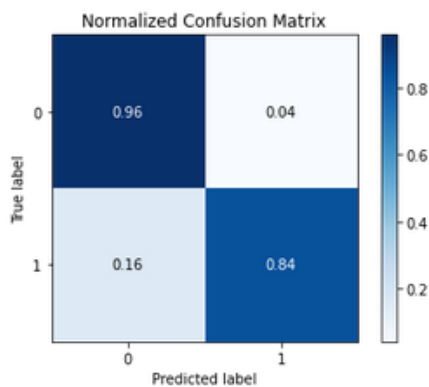


Figura 10 – Matriz de Confusão do Modelo Atual

POC I, acerca de variáveis como gênero e idade, e seu impacto na predição. Para isto, foi realizada uma análise dos resultados preditos de forma errada, para verificar se alguma das características estavam mais propensas a serem preditas de forma errada. Todavia, os resultados não apresentaram nenhuma implicação que pudesse ser extrapolada para seu uso em features.

## 5 Resultados Obtidos

A entrega do resultado se dará pela forma representada na Figura 11. Será realizado a predição e escrita da probabilidade de cada proteína ser positiva ou não. Esperava-se conseguir, também, uma acurácia mínima de 90% na classificação das proteínas isoladas.



Figura 11 – Resultado de um Exame de Imunofixação, positivo em G, K e L

Todavia, conforme mostrado nos resultados obtidos, notamos que a predição atual supera a marca que foi estabelecida como mínimo, bem como ultrapassa o que foi previsto, com detalhes importantes sobre a classificação do problema.

A homologação do sistema foi realizada no ambiente do laboratório. Como output, o modelo entrega um JSON, que é impresso no modelo de resultado pronto do Hermes Pardini. Esta etapa foi importante para nos insentarmos de qualquer dano causado por um erro de diagnóstico, uma vez que o modelo, nem o profissional que realiza a construção do mesmo, têm formação suficiente para afirmar diagnóstico. Portanto, o JSON gerado foi classificado como sugestão de diagnóstico, que pode ou não ser validado por um outro profissional. Na figura 12 temos um exemplo do output final, com ID do exame, classificação geral, quais proteínas são classificadas como positivas e o texto, que pode ser aplicado no resultado do exame.

Conforme a evolução do modelo foi progredindo, nota-se um avanço em sua acurácia. A figura 13 mostra que houveram variações no acerto do mesmo, uma vez que o projeto estava sujeiro a adição de novas imagens e evolução constante dos métodos para predição. O maior salto que tivemos, foi a utilização do método das regiões de interesse como validador, apresentando um salto significativo na acurácia.

```

{
  1.   "192839": {
  2.     "class": "positive",
  3.     "protein": ["G","K"],
  4.     "text": Padrão Biclonal Gamma Kappa
  5.   }
  6. }

```

Figura 12 – Exemplo de JSON passado para o sistema da Hermes Pardini

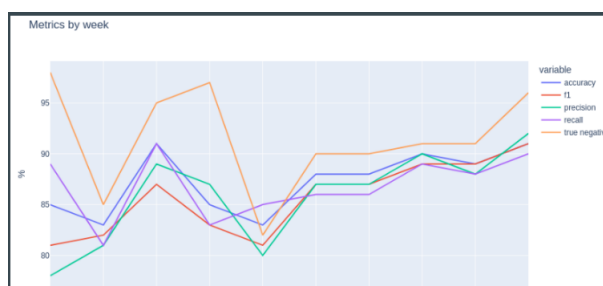


Figura 13 – Avanços mensais das métricas do modelo

Como o modelo retorna o grau de certeza de classificação da proteína, foi realizado um cálculo para compreender qual o limiar que o mesmo entende que um resultado é duvidoso. Ao analisar a classificação das proteínas que caíram nesse espaço de incerteza, obtivemos a figura 14. Nela, podemos ver que grande parte do erro foram de resultados falsos positivos, que não apresenta um prejuízo significativo para o laboratório, uma vez que a próxima etapa não é cara, e o paciente não tem risco de ir para casa com um exame negativo, mesmo tendo a doença.

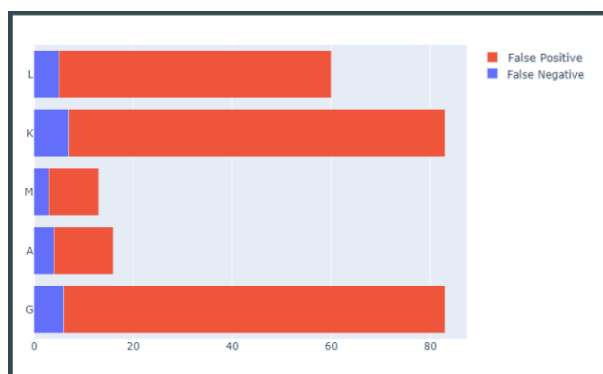


Figura 14 – Análise de erro do modelo

No quesito de facilitar a análise dos resultados, foi criada uma aplicação web, na qual o profissional faz upload do exame e o sistema realiza uma predição. A aplicação foi realizada para fins de validação, portanto não foi realizada nenhum teste unitário, nem prevenção de erros, uma vez que este não era o objetivo do trabalho. Na tela, o profissional

faz upload de um exame e o modelo retorna a classificação do modelo por meio do grau de certeza que o mesmo tem da proteína ser positiva.

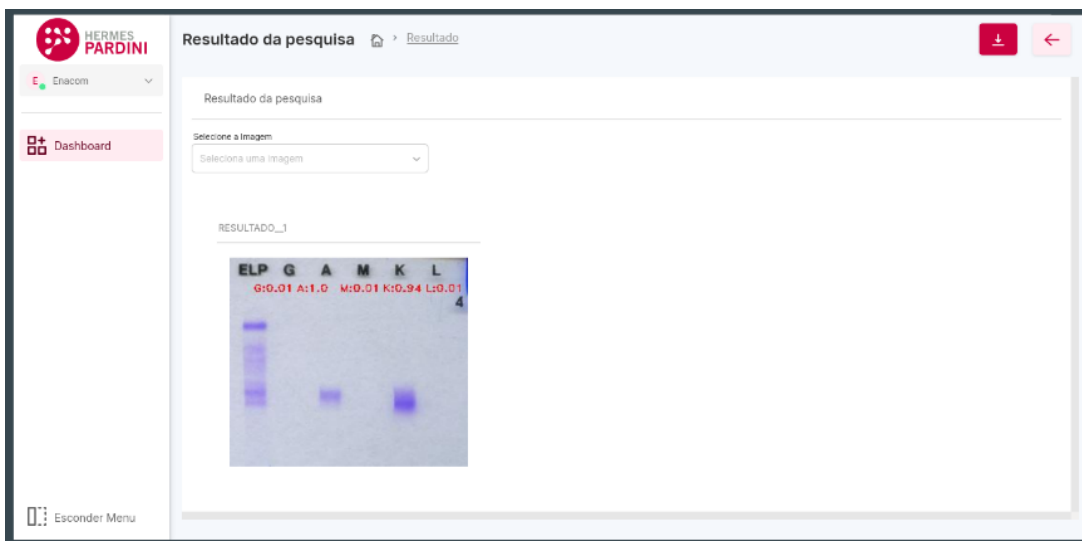


Figura 15 – Exemplo de tela do modelo de classificação

Por fim, com a soma de todos os métodos utilizados, chegamos ao resultado expressivo, mostrado pela matriz de confusão na figura 16, também apresentada anteriormente. Com isso, houve uma melhora significativa na acurácia, que foi para 91%. Não obstante, o acerto de exames negativos é de 96%, resultado muito satisfatório para o propósito do projeto.

Acurácia	F1	Precisão	Recall	True Negative
91%	91%	92%	90%	96%

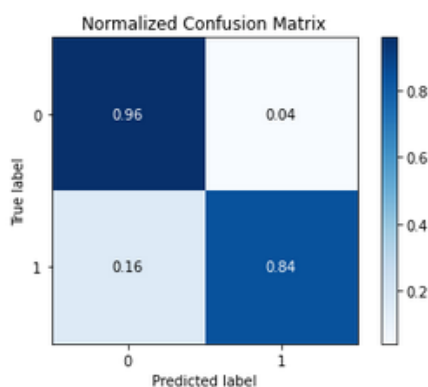


Figura 16 – Matriz de Confusão do Modelo Atual



## 6 Conclusão

O estudo realizado no laboratório Hermes Pardini demonstrou que a automatização do processo diagnóstico traz benefícios significativos, como a obtenção de resultados confiáveis em tempo mínimo. A implementação de um programa de automatização reduziu o tempo de laudo de exames de imunofixação de cinco a dez minutos para apenas alguns segundos. Isso proporciona uma maior rapidez na entrega dos resultados aos pacientes e aumenta a capacidade de atendimento do laboratório, processando um maior número de exames em menos tempo.

Além disso, a automatização do processo diagnóstico garante uma padronização e consistência nas análises, minimizando erros humanos e aumentando a precisão dos resultados. A utilização de algoritmos e inteligência artificial possibilita a realização de análises mais complexas e a identificação de padrões relevantes para um diagnóstico preciso.

O desenvolvimento deste trabalho representa um avanço significativo na área da saúde, proporcionando uma solução efetiva para agilizar o processo diagnóstico e entregar resultados confiáveis em tempo mínimo. A aplicação da tecnologia no laboratório Hermes Pardini demonstra o potencial da automatização na melhoria da eficiência e qualidade dos serviços prestados, contribuindo para o bem-estar dos pacientes e o aprimoramento da medicina diagnóstica como um todo.

Por fim, é importante ressaltar que a explicabilidade do modelo utilizado foi considerada durante o desenvolvimento do trabalho. Foram realizadas reuniões e alinhamentos periódicos com os usuários finais, visando garantir que os profissionais de saúde compreendessem, em linhas gerais, o funcionamento do sistema. Essa abordagem promoveu uma melhor aceitação por parte dos usuários e contribuiu para a utilização efetiva da tecnologia no processo diagnóstico.

Em suma, a automatização do processo diagnóstico, por meio do desenvolvimento deste trabalho, proporcionou benefícios como rapidez, confiabilidade e eficiência na obtenção de resultados, contribuindo para aprimorar a qualidade dos serviços prestados pelo laboratório Hermes Pardini e impactando positivamente a área da medicina diagnóstica.

# Referências

- [Ebrahimzadeh et al. 2018]EBRAHIMZADEH, E. et al. Prediction of paroxysmal atrial fibrillation: A machine learning based approach using combined feature vector and mixture of expert classification on hrv signal. *Computer Methods and Programs in Biomedicine*, v. 165, p. 53–67, 2018. ISSN 0169-2607. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0169260718304851>>.
- [Fluss, Faraggi e Reiser 2005]FLUSS, R.; FARAGGI, D.; REISER, B. Estimation of the youden index and its associated cutoff point. *Biometrical Journal*, v. 47, n. 4, p. 458–472, 2005. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.200410135>>.
- [Malek et al. 2020]MALEK, E. et al. Development of a machine learning algorithm for rapid, point-of-care prediction of serum monoclonal proteins in multiple myeloma. *Blood*, v. 136, p. 13–15, 2020. ISSN 0006-4971. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0006497118707738>>.
- [Wei et al. 2021]WEI, X.-Y. et al. Deep collocative learning for immunofixation electrophoresis image analysis. *IEEE Transactions on Medical Imaging*, v. 40, n. 7, p. 1898–1910, 2021.
- [Zhao et al. 2018]ZHAO, D. et al. Synthetic medical images using famp;rgan for improved lung nodules classification by multi-scale vgg16. *Symmetry*, v. 10, n. 10, 2018. ISSN 2073-8994. Disponível em: <<https://www.mdpi.com/2073-8994/10/10/519>>.