CECÍLIA APARECIDA SANTOS SILVA

USO DE CURVAS PRINCIPAIS E CONCEITOS DE REDES SIAMESAS NA CLASSIFICAÇÃO DE TUBERCULOSE EM FOTOGRAFIAS DE RADIOGRAFIA DE TÓRAX

LAVRAS – MG
2023

CECÍLIA APARECIDA SANTOS SILVA

USO DE CURVAS PRINCIPAIS E CONCEITOS DE REDES SIAMESAS NA CLASSIFICAÇÃO DE TUBERCULOSE EM FOTOGRAFIAS DE RADIOGRAFIA DE TÓRAX

Trabalho de conclusão de curso, no formato artigo, apresentado à Universidade Federal de Lavras, como parte das exigências do curso de Engenharia de Controle e Automação, para a obtenção do título de Bacharel.

Prof. Dr. Danton Diego Ferreira
Orientador

LAVRAS – MG

2023

CECÍLIA APARECIDA SANTOS SILVA

USO DE CURVAS PRINCIPAIS E CONCEITOS DE REDES SIAMESAS NA CLASSIFICAÇÃO DE TUBERCULOSE EM FOTOGRAFIAS DE RADIOGRAFIA DE TÓRAX
USE OF PRINCIPAL CURVES AND SIAMESE NETWORK CONCEPTS IN THE CLASSIFICATION OF TUBERCULOSIS IN CHEST X-RAY PHOTOGRAPHS

Trabalho de conclusão de curso, no formato artigo, apresentado à Universidade Federal de Lavras, como parte das exigências do curso de Engenharia de Controle e Automação, para a obtenção do título de Bacharel.

APROVADA em 11 de Dezembro de 2023.

Prof. Dr. Bruno Henrique Groenner Barbosa UFLA
Me. Fernando Elias De Melo Borges UFLA

Prof. Dr. Danton Diego Ferreira
Orientador

LAVRAS – MG

2023

AGRADECIMENTOS

Gostaria de agradecer à minha mãe, por ter plantado em mim o desejo de prosperar, aos meus tios, Aladir e Neide, por ter me possibilitado um ensino de qualidade, ao meu orientador Danton, pois as melhores oportunidades acadêmicas que surgiram vieram dele. Gostaria de agradecer também ao meu namorado, por toda a ajuda nesses 5 anos de faculdade e aos meus familiares e amigos pelo apoio.

RESUMO

Este trabalho propõe uma solução de *machine learning* para classificar imagens de raios-X do tórax para diagnosticar a tuberculose, uma das principais causas de morte por infecção no mundo, superada apenas pela covid-19 e de difícil detecção, especialmente em países em desenvolvimento e de grande extensão territorial como o Brasil.
A solução utiliza algoritmos baseados em curvas principais e redes neurais siamesas, que se mostram simples e eficazes, sem depender de redes neurais convolucionais (CNN), que são mais complexas e exigem mais recursos computacionais e dados. A base de dados utilizada foi a TBX11K, balanceada com 1600 imagens divididas em classe Tuberculose e Não-Tuberculose. O melhor modelo obtido na validação cruzada alcançou valores superiores a 80% em todas as métricas avaliadas, se aproximando dos resultados de outros trabalhos que usaram a CNN. A solução proposta pode ser útil para auxiliar os profissionais de saúde na triagem e na detecção precoce da tuberculose, reduzindo o risco de uma triagem inadequada e o surgimento de tuberculose multirresistente.

Palavras Chaves: Curvas Principais, Redes Neurais Siamesas, Tuberculose, Machine Learning, Redes Neurais Convolucionais.

ABSTRACT

This work proposes a machine learning solution to classify chest X-ray images to diagnose tuberculosis, one of the main causes of death from infection in the world, surpassed only by covid-19 and difficult to detect, especially in developing countries and with a large territorial extension like Brazil.
The solution uses algorithms based on principal curves and Siamese neural networks, which prove to be simple and effective, without relying on convolutional neural networks (CNN), which are more complex and require more computational resources and data. The database used was TBX11K, balanced with 1600 images divided into Tuberculosis and Non-Tuberculosis classes. The best model obtained in cross-validation achieved values above 80% in all metrics evaluated, approaching the results of other studies that used CNN. The proposed solution could be useful in assisting healthcare professionals in the screening and early detection of tuberculosis, reducing the risk of inadequate screening and the emergence of multidrug-resistant tuberculosis.

Keywords: Principal Curves, Siamese Neural Networks, Tuberculosis, Machine Learning, Convolutional Neural Network.

SUMÁRIO

INTRODUÇÃO

A tuberculose é uma das doenças infecciosas mais mortais do mundo, ficando atrás apenas da covid-19. Em 2022, 1,3 milhão morreram de tuberculose enquanto 10,6 milhões adoeceram, embora essa seja uma doença curável e evitável. Ela é causada pela bactéria Mycobacterium tuberculosis e afeta geralmente os pulmões, mas pode acometer qualquer outro órgão. Nem todas as pessoas infectadas desenvolvem o que é chamada de tuberculose ativa, principalmente os contatos de pacientes infectados correm o risco de contrair a bactéria e possuírem a forma latente da doença (WORLD HEALTH ORGANIZATION, 2019). Essas pessoas totalizam cerca de 25% da população mundial, sendo que nos 2 primeiros anos após a contaminação, é elevadíssimo o risco da tuberculose latente evoluir para sua forma ativa (MINISTÉRIO DA SAÚDE, 2019), o que torna de suma importância tratar a tuberculose latente no combate a tuberculose ativa (WORLD HEALTH ORGANIZATION, 2018).

Cada forma de tuberculose é tratada com medicamentos específicos. Quando o tratamento é realizado de maneira inadequada as bactérias não respondem aos remédios, dando origem a tuberculose multirresistente, limitando ainda mais as opções terapêuticas. Por isso, o Ministério da Saúde do Brasil recomenda a exclusão de tuberculose ativa por meio da triagem de sintomas e a radiografia de tórax, para todos os contatos de pacientes com TB pulmonar ativa antes de iniciar o tratamento preventivo (MINISTÉRIO DA SAÚDE, 2019).

Além da complexidade em diagnosticar a doença, a qualidade das imagens de radiografia e a capacidade do profissional da saúde responsável por realizar o exame médico, são fatores decisórios no resultado. Dado o crescente papel da tecnologia em diversas áreas, trazendo bons desempenhos e assertibilidade, como na predição de incidência de casos de malária (BARBOZA, 2021) e identificação de grau de risco de pé diabético (RESENDE, 2021), a OMS passou a recomendar o uso de diagnósticos auxiliados com computadores em RxT para o rastreamento de tuberculose (WORLD HEALTH ORGANIZATION, 2023).

Para classificação de imagens, é comum o uso de redes neurais convolucionais (*convolutional neural networks*, CNN). Contudo, elas possuem muitos pesos a serem treinados, carecendo de uma quantidade massiva de imagens, agregando complexidade computacional ao modelo, o que restringe os meios pelos quais elas podem ser utilizadas (PONTI, 2018). Somado a isso, muitas regiões de países subdesenvolvidos vivem em situações precárias quando o assunto é acesso à saúde de qualidade, sem equipamentos que sustentem algoritmos pesados.

A partir disso, surge a preocupação de desenvolver algoritmos eficientes que podem ser executados em qualquer smartphone, por exemplo. Pensando nisso, o presente trabalho discute a implementação de técnicas clássicas, como as redes neurais siamesas associadas à utilização de curvas principais como seletoras de características no desenvolvimento de um modelos que além de serem mais leves comparado às CNNs apresentam resultados competitivos.

As curvas principais são uma generalização não linear de análises de componentes principais e vêm sendo usadas em aplicação de machine learning para classificação devido sua baixa complexidade computacional (HASTIE, 1989). Quando se trata de problemas com poucos dados, em análises de imagens, as redes neurais siamesas apresentam bons desempenhos, sendo uma arquitetura de rede que compara a saída de duas CNNs, as quais compartilham os mesmos pesos, sendo que uma das CNNs é tida como referência e a outra processa o objeto a ser classificado. Esses fatores foram decisivos para a escolha desses algoritmos para compor a solução substituta ao uso das CNNs.

Uso de curvas principais e conceitos de redes siamesas na classificação de tuberculose via fotografias de radiografia de toráx

1st Cecília Aparecida Santos Silva
Universidade Federal de Lavras
Lavras, Brasil
cecilia.silva1@estudante.ufla.br

2nd Danton Diego Ferreira
Universidade Federal de Lavras
Lavras, Brasil
danton@ufla.br

***Resumo*—A tuberculose é uma das doenças que mais matam no mundo, e o tratamento preventivo desempenha um papel fundamental no combate a essa doença. Para isso, é necessário tratar a infecção latente por *Mycobaterium tuberculosis* (ILTB), já que ela afeta cerca de 25% da população mundial. No Brasil é recomendado a radiografia de tórax (RxT) nos contatos de pacientes com TB pulmonar ativa antes de iniciar o tratamento preventivo para se certificar que o paciente não tem tuberculose ativa. No entanto, a dificuldade de identificar tuberculose ativa em radiografias de tórax e a necessidade de profissionais experientes impactam nesse processo. Ao mesmo tempo, nota-se o crescimento do uso de inteligência artificial como apoio aos profissionais de saúde, muitas vezes usando algoritmos de *deep learning*, que são computacionalmente complexos e carecem de muitos dados para seu treinamento. O objetivo deste trabalho é verificar a possibilidade de usar curvas principais e redes neurais siamesas como uma alternativa para classificação de Tuberculose em fotografias de radiografias de tórax. Utilizou-se uma base de dados balanceada com 1600 imagens divididas em classe Tuberculose e Não- Tuberculose. Os resultados de desempenho da abordagem proposta na validação cruzada foram superiores a 70%.**

***Index Terms*—curvas principais, redes neurais siamesas, tuberculose**

I. Introdução

Todos os anos, no mundo, cerca de 1,5 milhões de pessoas morrem vítimas da tuberculose [2]. Na Figura 1 são mostradas as taxas de mortalidade e acometimento de tuberculose de 2009 a 2020, segundo dados coletados dos *reports* de tuberculose global disponibilizados pela organização mundial da saúde [1]. Como agravante, os contatos de pacientes infectados correm o risco de contrair o *Mycobacterium tuberculosis*, que, na maioria dos casos, torna-se uma infecção latente por ser contido pelo sistema imune. No entanto, principalmente nos dois primeiros anos após a contaminação, essa infecção latente pode evoluir e se tornar tuberculose ativa [3]. Dessa forma, tratar a infecção latente desempenha um papel importante no combate à tuberculose ativa [4].

Dada a diferença entre o tratamento da tuberculose ativa e latente, é necessária a exclusão da tuberculose (TB) ativa por meio da triagem de sintomas e a radiografia de tórax (RxT), conforme recomendação da OMS [4] e do Ministério da Saúde

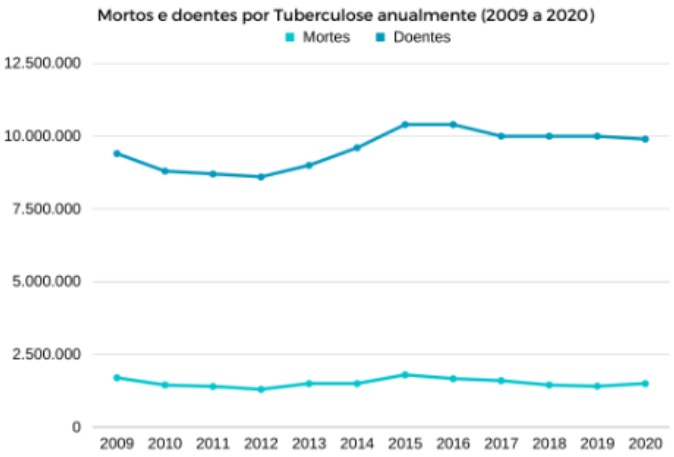

Figura 1. Quantidade de mortes e adoecidos por tuberculose entre 2009 e 2020
Fonte: Global Tuberculosis reports [1].

do Brasil [3]. A RxT é recomendada, no Brasil, para todos os contatos de pacientes com TB pulmonar ativa antes de iniciar o tratamento preventivo [3]. No entanto, a falta da RxT e a necessidade de especialistas para interpretar os resultados atrasam ou impedem o processo e, consequentemente, o início do tratamento.

Paralelo a isso, a Organização Mundial da Saúde passou a recomendar o uso de diagnósticos auxiliados por computador, do ingles *computer-aided diagnosis* (CAD), que são ferramentas que auxiliam na detecção e classificação de doenças, em RxT para o rastreamento de tuberculose [5], a fim de ajudar o profissional de saúde na tomada de decisão.

Em CADs, principalmente quando são desenvolvidos para análise de imagens, usam-se algoritmos de *deep learning*, a exemplo do auxílio ao diagnóstico de COVID-19 e outras pneumonias em imagens de radiografias de tórax (RxT) [6], bem como do aprendizado não supervisionado de classificadores especialistas para apoiar a triagem de Covid-19 com base em dados de tomografia computadorizada [7]. Esses algoritmos usam muitas camadas de processamento de informações

não lineares de natureza hierárquica, o que leva à necessidade de quantidades massivas de dados rotulados para que a rede possa aprender conceitos simples [8]. Relacionado a isso, tem a complexidade para encontrar os melhores parâmetros e também seu custo computacional.

Por outro lado, tem-se as curvas principais, que são uma generalização não linear de análises de componentes principais e vêm sendo usadas em aplicações de *machine learning* para classificação devido à sua baixa complexidade computacional na fase operacional. Além disso, quando se trata de problemas com poucos dados, em análises de imagens, as redes neurais siamesas apresentam bons desempenhos. Devido a esses fatos e aos contrapontos dos algoritmos de *deep learning*, surge a oportunidade de usar tais ferramentas em problemas de classificação de imagens como é o caso da triagem de tuberculose. Há trabalhos referentes ao uso de curvas principais para esse propósito, porém são usadas as redes neurais convolucionais (CNN) como redutor de dimensionalidade, como é o caso do trabalho que trata da utilização de curvas principais na triagem de pacientes com tuberculose [9]. O presente trabalho consiste em utilizar curvas principais como seletoras de características e o uso de conceitos de redes neurais siamesas para a classificação de tuberculose em fotografias de radiografias de tórax, numa tentativa de obter bons resultados sem o uso de *deep learning*.

II. Análise de componentes principais

A análise de componentes principais (PCA) é uma ferramenta estatística de transformação ortogonal dos eixos de coordenadas de um sistema multivariado. Ela converte um conjunto de dados provavelmente correlacionados em um conjunto de valores linearmente não correlacionados. A PCA é amplamente utilizada nas áreas de reconhecimento de padrões e visão computacional para extração de características e redução de dados multidimensionais a dimensões inferiores. Isso é feito ao mesmo tempo que retém a maior parte das informações, com base no número de componentes desejadas ou variância estipulada [10].

III. Curvas principais

As curvas principais são uma generalização não linear do PCA. Elas foram definidas [11] como curvas parametrizadas unidimensionais com propriedade de autoconsistência que passam por um conjunto de dados no espaço multidimensional descritos por X no espaço de dados original, fornecendo uma boa representação em uma dimensão. Elas têm sido usadas para representação de dados e análise dado sua boa capacidade de representação dos dados, ao capturar sua não linearidade multidimensional, retomando a distribuição dos mesmos unidimensionalmente [12].

Uma curva principal é um vetor *f(t)* em *d* funções contínuas e uma única variável *t*, ou seja, $f(t) = [f_1(t), f_2(t), ..., f_d(t)]^T$. Essas funções são chamadas funções de coordenadas e o parâmetro *t* está relacionado a ordenação ao longo da curva.

Seja f uma curva suave no intervalo fechado $I \subseteq \mathbb{R}^1$ que não se intercepta, ou seja, $t_1 \neq t_2 \rightarrow f(t_1) \neq f(t_2)$ e com seu comprimento finito dentro de um esfera de dimensões finitas em $\mathbb{R}^d$.

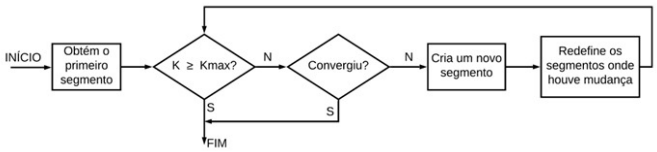

Figura 2. Diagrama em blocos do algortimo k-segments
Fonte: Uso de Curvas Principais em Seleção Inteligente de Dados para ambientes com alta taxa de eventos [15].

Tomando x como um vetor aleatório em $\mathbb{R}^d$, o índice de projeção $t_f : \mathbb{R}^d \rightarrow \mathbb{R}^1$ é definido como:

$$t_f(x) = sup\{t : \|x - f(t)\| = inf\|x - f(\mu)\|\} \qquad (1)$$

onde $\|.\|$ refere-se a a norma Euclidiana em $\mathbb{R}^d$ e μ é uma variável auxiliar definida em $\mathbb{R}^1$. A projeção índice $t_f(x)$ é o valor de *t* para o qual a curva principal *f(t)* é mais perto de *x*. Se houver mais de um valor possível, o maior valor é selecionado. Assim, os pontos que compõem a curvas principais são a média dos dados projetados nela [11], como:

$$f(t) = E[x|t_f(x) = t]\forall t \qquad (2)$$

A. K-segments

As curvas principais foram extraídas usando o algoritmo de k-segmentos proposto por Verbeek [13] e diponibilizado em versão Python pelos autores do trabalho apresentado em [14]. Esse algoritmo, primeiramente, localiza k segmentos de linhas distintos no conjunto de dados. Após isso, esses segmentos são vinculados uns aos outros para compor uma linha poligonal, que pode ser usada como uma primeira curva principal.

Para construir uma curva principal satisfatória, deve-se ajustar principalmente o parâmetro de k-segmentos, que representa a quantidade de segmentos que formam a curva principal. Esse valor deve ser suficientemente grande para refletir a complexidade dos dados e pequeno o suficiente para não sobreajustar aos dados de treinamento [14]. Há também o parâmetro λ, que, segundo [16], determina a correta ligação entre os segmentos da curva. Principalmente em dados que apresentam interseções, o efeito de sua modificação é notável. Em [13], são definidos que valores de λ de $1/2$ a 2 apresentam bons resultados para suavizar a curva. A Figura 2 mostra o diagrama do algoritmo k-segments para a obtenção da curva principal. Ja na Figura 3 é ilustrada uma curva gerada pelo algortimo.

IV. Redes Neurais Siamesas

A configuração original das redes neurais siamesas consiste em dois redes neurais convolucionais (CNN) que compartilham os mesmos pesos [17], conforme ilustrado na Figura 4. Uma das CNNs processa uma imagem de referência, enquanto a outra processa as demais imagens. Como saída, tem-se dois vetores de características para cada amostra analisada, que

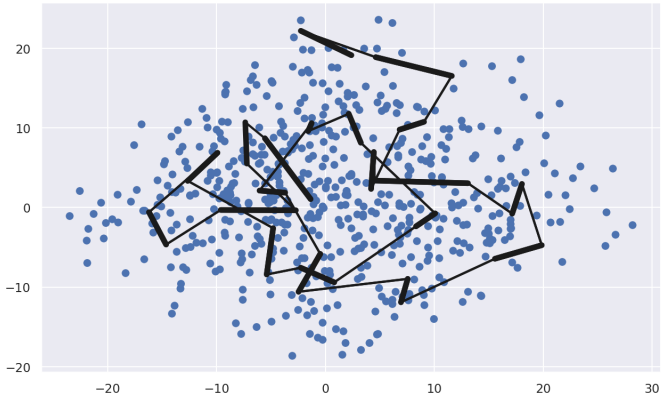

Figura 3. Exemplo de uma curva principal

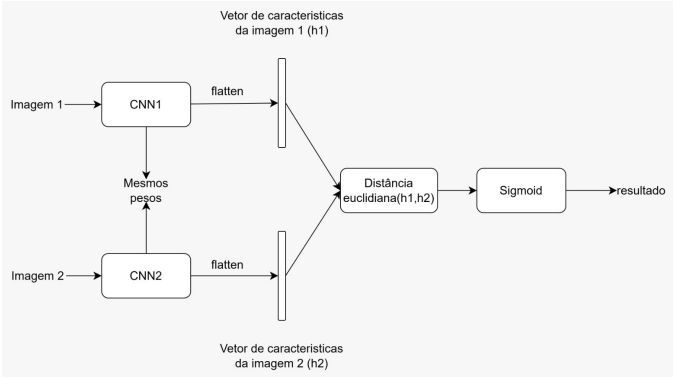

Figura 4. Arquitetura original da rede neural siamesa
Fonte: Do autor.

serão comparados pelo algoritmo por meio de uma medida de distância. O vetor resultante dessa comparação pode então ser processado por uma função de ativação sigmoidal para gerar a classificação.

As duas métricas de distância mais usadas são: (i) distância euclidiana, que é a distância entre dois pontos que pode ser provada pela aplicação repetida do teorema de Pitágoras; (ii) a distância de Manhattan, em que a distância entre dois pontos é a soma das diferenças absolutas de suas coordenadas; (iii) e a similaridade de cosseno, que mede a similaridade entre dois vetores num espaço vetorial calculando o cosseno do ângulo formado por eles [18].

V. XGBoost

O XGBoost é um algoritmo de *machine learning* baseado em *Gradient Boosting Machine* e é amplamente utilizado em problemas de classificação e regressão devido à sua eficiência e flexibilidade [19].

O *Gradient Boosting Machine* (GBM) é um tipo de *ensemble* cuja proposta é combinar iterativamente várias árvores de decisão tidas como fracas para obter um modelo com desempenho desejável. Esse método pode ser interpretado como um método de otimização que visa encontrar um modelo aditivo, abordagem estatística para extrair uma explicação das variáveis dos preditores lineares [20]. O algoritmo GBM adiciona iterativamente, em cada etapa, uma nova árvore de decisão (ou seja, ”aluno fraco”) que melhor reduz a função de perda. Esse processo continua até que um número máximo de iterações, fornecido pelo usuário, seja atingido. Ao ajustar árvores de decisão aos resíduos, o modelo é melhorado nas regiões onde não apresenta bom desempenho [21].

Alguns dos hiperparâmetros desse modelo foram analisados no trabalho em questão. A taxa de aprendizado, também chamada de *learning rate*, está relacionada com a quantidade de pequenos passos dados pelo modelo no processo de treinamento e influencia diretamente em sua à medida que vai diminuindo, geralmente entre 0 e 1. Porém, esse parâmetro é inversamente proporcional ao número de iterações [21].

O hiperparâmetro *random-state* também melhora as métricas do modelo, além de reduzir o custo computacional do algoritmo por um fator equivalente ao fator de subamostragem. Em cada etapa iterativa, em vez de usar o conjunto de dados de treinamento completo, é usada uma subamostra selecionada aleatoriamente. Ressalta-se a necessidade de verificar vários valores de uma fração da subamostra para avaliar o impacto da diminuição do número de pontos de dados na qualidade do ajuste do modelo [21]. Já *max-depth* está relacionado à profundidade da árvore e deve ser selecionado considerando que, quanto maior, mais complexo, computacionalmente custoso e propenso a sobreajuste. O *subsample* também evita sobreajuste adicionando aleatoriedade às amostras de cada iteração [19].

VI. Arquitetura de solução

No presente trabalho, na fase operacional, as imagens serão vetorizadas, reduzidas com redimensionamento e também com a aplicação da Análise de Componentes Principais (PCA), que ao mesmo tempo agirá como um seletor de características. Na próxima fase, a curva principal será gerada. Será usado o conceito de redes neurais siamesas relacionado à medida de distância euclidiana entre uma amostra referência e as demais. As CNNs da siamesa serão substituídas por curvas principais, formando portanto o conceito de curvas principais siamesas. As saídas de cada curva principal que compõe a nova estrutura da siamesa será o vetor de mapeamento da entrada nas curvas principais. Neste caso, o vetor será composto pela distância da entrada (imagem vetorizada e pré-processada com PCA) aos segmentos da curva principal. Além disso, no lugar da função de ativação sigmoidal, estará o XGBoost. Na Figura 5 o processo é ilustrado.

VII. Base de dados

A base de dados usada neste trabalho é a TBX11K, a qual foi construída especialmente para auxiliar no diagnóstico de TB devido à alta qualidade de suas amostras [22]. Ela contém 11.200 imagens de raios-X com um tamanho de 512x512, divididas em três categorias: Saudável, Doente (mas Não TB), e TB, além de imagens de TB incertas, que foram colocadas no grupo de imagens de teste, e outros *datasets*. Apesar disso, há apenas 800 imagens de TB, enquanto as outras duas categorias possuem 3.800 cada, o que evidencia um grande

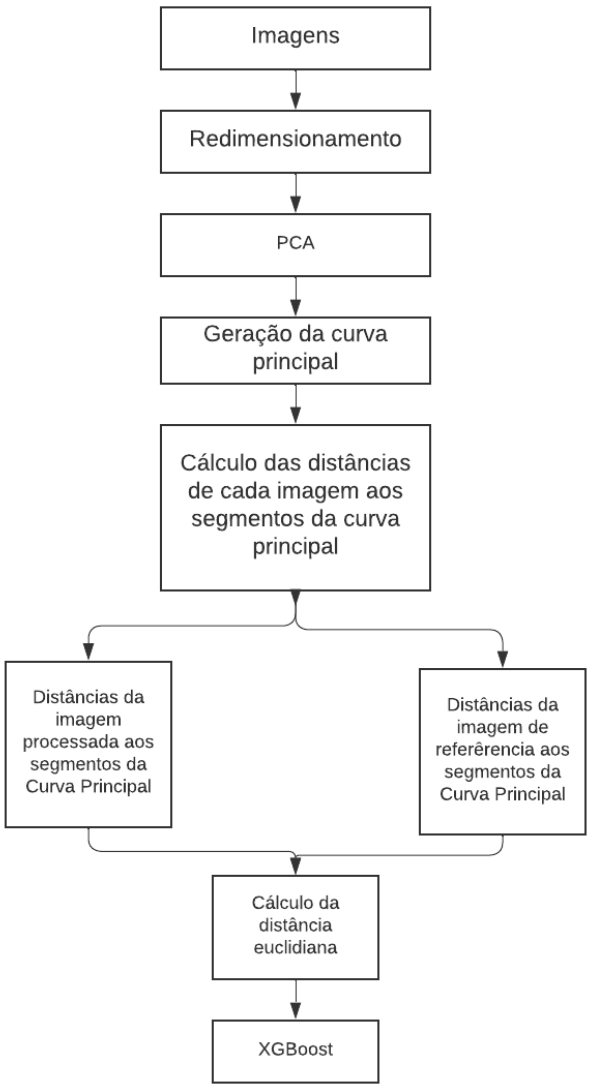

Figura 5. Arquitetura do processo

desbalanceamento. Para este trabalho, apenas imagens das três categorias foram utilizadas, e houve um balanceamento a fim de trabalhar com 800 imagens de TB e 800 imagens de não TB.

Foram selecionadas aleatoriamente 800 imagens das categorias Saudável e Doente, mas não TB, para compor a classe Não-TB, tendo a mesma quantidade de imagens que a classe TB. Essa atitude é importante para evitar que o modelo se torne enviesado e impreciso, impactando negativamente a classe minoritária, que, nesse caso, seria a classe de TB, o que é problemático por se tratar de um diagnóstico médico [23]. Na Tabela I é indicado o resultado do balanceamento.

VIII. Método proposto

Como uma primeira estratégia de redução de características, as imagens foram redimensionadas para 128x128 pixels.

Tabela I
Quantidade de Imagens por Classe

Classe	Categoria	Quantidade de Imagens
TB	TB	800
Não - TB	Saudável	422
	Doente mas não TB	378

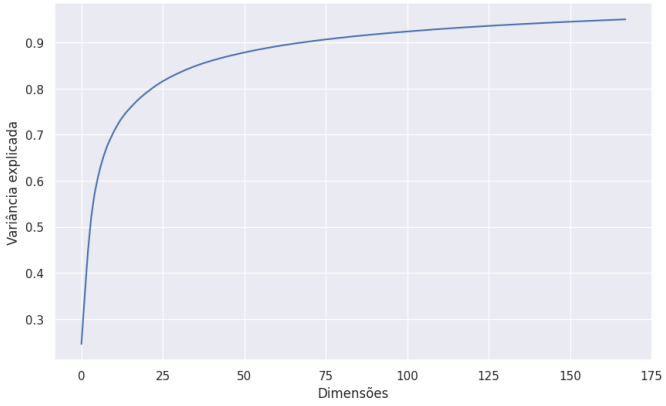

Figura 6. Gráfico de energia do PCA

Houve também a conversão de RGB para o padrão de cores gray, de forma a tornar suas representações bidimensionais. Cada pixel possui um valor entre 0 e 255, onde 0 representa preto e 255 representa branco. Devido a isso, também foi realizada uma normalização, dividindo todos os pontos por 255, de forma a deixar os valores entre 0 e 1, uma vez que usa-se distância euclidiana nas redes neurais siamesas.

Com o auxílio do método *flatten* da biblioteca *numpy* do Python, que coloca uma linha na frente da outra, foi realizada a vetorização dos dados, ou seja, cada imagem passou a ser representada por um vetor de 16384 características, já que tanto o PCA quanto o k-segments devem ser imputados com uma matriz bidimensional onde as linhas representam as amostras. É perceptível que esse número é muito elevado ao levar em conta o intuito de obter um método de baixo custo computacional. A fim de reduzir a dimensão e eliminar redundâncias, os dados foram processados pelo PCA que foi configurado para manter uma variância de 95% dos dados originais. Para isso, a base de dados foi dividida em 70% treino e 30% teste, onde a base de treino foi usada para treinar o pca resultando em 168 componentes. O gráfico de energia do PCA é apresentado na Figura 6.

Há dois grupos de parâmetros a serem selecionados, aqueles que resultam nas melhores curvas principais e aqueles que resultam no melhor classificador possível. Uma vez que eles estão relacionados, a otimização é realizada de maneira aninhada. Os dados de treino geram a curva principal que representa a TB, usando apenas a base de dados da classe TB, portanto, a curva resultante é representativa da classe TB. Em seguida, as imagens de TB do banco de dados de treino são mapeadas na curva, ou seja, suas distâncias aos segmentos da curva são obtidas. A imagem que possui a menor distância

é fixada como referência na entrada de uma das gêmeas. A partir de então, os dados de entrada do XGBoost são as distâncias euclidianas de cada imagem àquela fixada, sendo que a quantidade de características corresponde à quantidade de segmentos da curva.

Os parâmetros testados das curvas principais foram combinados e armazenados em uma matriz a ser percorrida por um laço de repetição. Para o XGBoost, foi usada a técnica de ajuste de hiperparâmetros *GridSearch*, que permite que o conjunto de valores para cada hiperparâmetro a ser investigado seja especificado e, em seguida, treine o modelo com cada combinação possível a fim de encontrar o melhor desempenho. Dessa forma, para cada linha da matriz, foi executado um *GridSearch*. A melhor acurácia de cada repetição foi armazenada em um dicionário onde a chave armazena o recall e ambos grupos de parâmetros analisados no momento, e o valor é a melhor acurácia encontrada. Na Tabela II encontram-se os parâmetros testados, bem como o conjunto de valores.

Tabela II
CONFIGURAÇÕES DE HIPERPARÂMETROS

Grupo de Parâmetros	Hiperparâmetro	Conjunto de Valores
Curvas principais	K	25-40, 5 em 5
	Lambda	0.5-2.0, 0.5 em 0.5
XGBoost	learning_rate	0.01 ; 0.05 ; 0.10
	max_depth	3 ; 6 ; 8 ; 10 ; 12
	max_leaves	2 ; 4 ; 6
	min_child_weight	4 ; 5 ; 6
	colsample_bytree	0.3 ; 0.4
	n_estimators	650 ; 1000 ; 1500
	objective	binary:logistic
	subsample	0.5 ; 1
	random_state	42

Os melhores parâmetros foram selecionados com base na melhor acurácia das iterações. O *recall* foi armazenado como critério de desempate. A análise foi realizada sobre a acurácia, já que a base está balanceada. Dito isso, o valor dessa métrica indica que o modelo acertou bem ambas as classes. Ademais, como se trata de um problema da área da saúde, é preferível que o modelo erre dizendo que pessoas saudáveis estão doentes do que o contrário, posto que é muito mais arriscado para o paciente ser classificado como saudável estando doente. Por esse motivo, usa-se o recall como segundo critério para definir os melhores parâmetros.

Em seguida, usou-se o *k-fold*, com k igual a 10, para validar o classificador. O *k-fold* é um tipo de validação cruzada (*cross-validation*) útil para avaliar a eficácia do modelo desenvolvido, que consiste em dividir o conjunto de dados em k partes, usando $k-1$ para treinamento e o restante para validação.

O método proposto é ilustrado no diagrama em blocos presente na Figura 7.

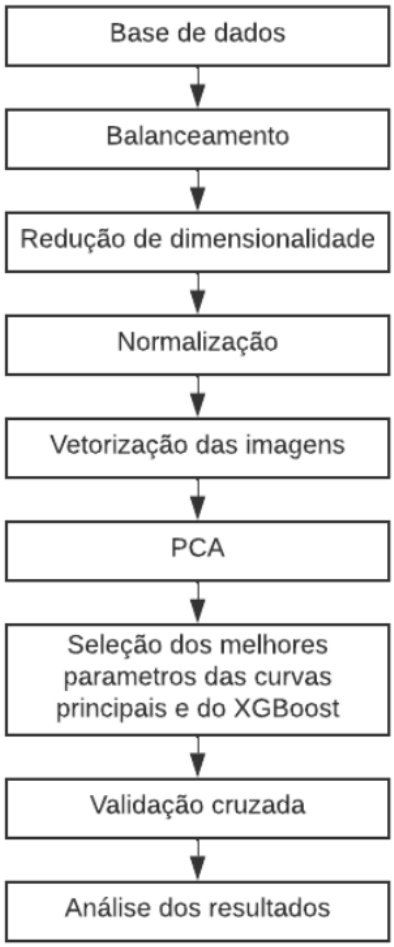

Figura 7. Fluxograma do desenvolvimento do método proposto.

IX. RESULTADOS

Devido o risco de classificar como Não-TB imagens de TB, surge a necessidade de analisar qual o valor de probabilidade será usada para determinar a classe, o que normalmente é 50%. Ao executar o modelo na etapa de otimização, percebeu-se que as probabilidades eram bem distribuidas e comportavam-se de maneira similar em todas as execuções. Na figura 8 é mostrado o boxplot da probabilidade de ser não-TB de uma das execuções do modelo. Essa figura ilustra quais as probabilidades resultantes do modelo para cada classe, sendo que a classe Não-TB está em vermelho e a classe de TB está em azul. Assim, nota-se que a maioria das imagens de Não-TB obtiveram probabilidade maior que 60%de ser Não-TB, enquanto a maioria das imagens de TB, em azul, receberam probabilidade menor que 40% para ser Não-TB. A partir dele, objetivou-se encontrar uma probabilidade que penalizasse o *recall* Não-TB para aumentar o *recall* TB, mas que o mantivesse próximo a 70%. Assim, escolheu-se a probabilidade de 65%, visto que, apenas 27,49% das imagens Não-TB obtiveram essa probabilidade.

Na Tabela III mostra-se os hiperparâmetros encontrados que geraram a melhor acurácia da etapa de otimização. Os mesmos são usados no modelo na etapa de validação cruzada. Na Tabela IV, tem-se as métricas geradas por eles, ressaltando que o *recall* da classe TB também foi o maior dentre os gerados na etapa de otimização.

A fim de visualizar a distribuição das amostras nas classes

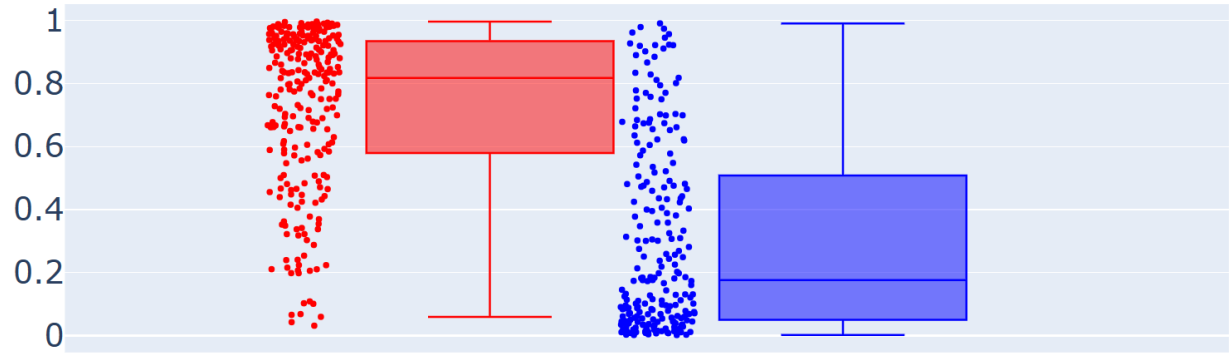

Figura 8. Boxplot da probabilidade de ser Não-TB. Vermelho representa as probabilidades de ser Não-TB para imagens Não-TB e Azul são as probabilidades de ser Não-TB para imagens de TB.

Tabela III
CONFIGURAÇÃO DOS MELHORES HIPERPARÂMETROS

Algoritmo	Hiperparâmetros	Valores
Curvas principais	Números de segmentos	45
	Lambda	1.5
XGBoost	Learning rate	0.05
	Max depth	3
	Max leaves	6
	Min child weight	4
	N estimator	1000
	Subsample	0.5
	Colsample_bytree	0.3

Tabela IV
RESULTADOS DAS MÉTRICAS DE DESEMPENHO

Métrica	Resultado
Recall TB	78.60%
Recall Não-TB	72.51%
Precisão TB	72.29%
Precisão Não-TB	78.79%
Acurácia	75.42%

previstas, na Tabela V é possível ver a matriz de confusão desses resultados. Nota-se que 49 imagens de TB frente a 69 imagens de Não-TB foram classificadas erroneamente.

Tabela V
TABELA DE CONFUSÃO DOS MELHORES HIPERPARÂMETROS

		Classe Prevista	
		Não-TB	TB
Classe Real	Não-TB	182	69
	TB	49	180

Os resultados da etapa de validação cruzada estão expressos a seguir. Os valores médios das métricas do *k-fold* estão na Tabela VI. Analisando-os, percebe-se que foram próximos àqueles obtidos na etapa de otimização. Além disso, os desvios padrões encontrados, por serem baixos, indicam que o modelo tem boa capacidade de generalização.

Tabela VI
RESULTADOS DAS MÉTRICAS DE DESEMPENHO COM DESVIO PADRÃO

Métrica	Resultado	Desvio Padrão
Recall TB	83.10 %	±4.77 %
Recall Não-TB	69.50 %	±7.32 %
Precisão TB	73.10 %	±8.01 %
Precisão Não-TB	80.60 %	±4.09 %
Acurácia	76.30 %	±4.40 %

Os resultados dos 10 *folds* estão na Tabela VII. Os mesmos reforçam a generalização do modelo. Nota-se que os folds 4, 7 e 10 são os responsáveis por aumentar o desvio padrão do recall Não-TB enquanto, que para a precisão TB, há uma variedade maior de valores, o que eleva o desvio. Para as métricas restantes, observa-se pouca discrepância entre os folds.

Levando em consideração a acurácia e o recall-TB, que garantem que, além do modelo acertar a previsão de ambas as classes com bom desempenho, ele também é bom em identificar imagens de TB, dentre as previsões de TB, o melhor *fold* foi o quarto. Na Tabela VIII evidencia-se suas métricas.

Na Tabela IX, tem-se a matriz de confusão do melhor modelo. A quantidade de imagens classificadas de maneira equivocada nas duas classes é bem equilibrada. Ademais, as métricas encontradas são superiores a 80%, indicando um bom desempenho desse *fold*.

Para avaliar o modelo, ressalta-se a não utilização de CNN em nenhum processo, fato importante visto que esses são modelos que melhor classificam imagens. Dito isso, consideram-se os resultados acima como satisfatórios, uma vez que não só apresentam capacidade de generalização, um *fold* com resultados superiores a 80%, mas também a possibilidade de se usar algoritmos clássicos em problemas de imagens, o que traz a vantagem de um custo computacional reduzido, visto que sua fase operacional é simples, já que para cada imagem, o que deve ser feito é, redimensionamento, vetorização, operação para reduzir dimensão com a matriz construída do PCA,

Tabela VII
RESULTADOS POR FOLD

Fold	Recall TB	Recall Não-TB	Precisão Não-TB	Precisão TB	Acurácia
1	80,00%	67,00%	83,00%	63,00%	73,00%
2	83,00%	70,00%	78,00%	76,00%	77,00%
3	85,00%	76,00%	76,00%	85,00%	81,00%
4	83,00%	86,00%	83,00%	86,00%	84,00%
5	86,00%	65,00%	83,00%	71,00%	76,00%
6	76,00%	70,00%	75,00%	71,00%	73,00%
7	90,00%	60,00%	86,00%	69,00%	75,00%
8	89,00%	71,00%	86,00%	76,00%	80,00%
9	83,00%	67,00%	79,00%	72,00%	75,00%
10	76,00%	63,00%	77,00%	62,00%	69,00%

Tabela VIII
RESULTADOS DAS MÉTRICAS DE DESEMPENHO DO MELHOR FOLD

Métrica	Resultado
Recall TB	82.93 %
Recall Não-TB	85.90 %
Precisão TB	86.08 %
Precisão Não-TB	82.72 %
Acurácia	84.38 %

Tabela IX
TABELA DE CONFUSÃO DO MELHOR FOLD

		Classe Prevista	
		Não-TB	TB
Classe Real	Não-TB	67	11
	TB	14	68

calculo de 45 distâncias euclidianas (45 segmentos), cálculo de 45 subtrações ao quadrado e finalmente aplicação no XGBoost.
.

A título de comparação, o trabalho [9] realiza a classificação usando curvas principais, no entanto, usando CNN como seletoras de características. Nele, o *recall-TB* foi de 84%, e a acurácia foi de 89%. Apesar de inferiores, os resultados do presente trabalho são competitivos.

O uso de CNN foi substituído pelo redimensionamento e vetorização das imagens. Diferente da CNN, essa estratégia não trabalha bem como seletora de características, o que impactou negativamente os resultados, mas diminuiu o custo computacional.

X. CONCLUSÃO

A tecnologia tem sido presente em diversas áreas, auxiliando nas tomadas de decisão. Principalmente na saúde, ferramentas de *machine learning* podem contribuir como uma triagem, apoiando os profissionais de saúde na priorização dos pacientes a serem atendidos, o que acontece devido às condições adversas que os mesmos enfrentam, em especial em países emergentes e de proporções continentais como o Brasil.

Devido ao impacto negativo de uma triagem errada, algoritmos de *machine learning* para saúde requerem um cuidado especial. Em contrapartida, essas mesmas condições adversas impedem o uso de algoritmos muito robustos. Ao mesmo tempo, há doenças como a tuberculose, que são extremamente preocupantes e difíceis de serem identificadas. Portanto, é imprescindível buscar uma solução que identifique bem a tuberculose, ao mesmo tempo que seja leve ao ponto de ser usada em um *smartphone*. Dessa maneira, esse trabalho buscou analisar a possibilidade de usar curvas principais, conceitos de redes neurais siamesas associadas ao XGBoost, para classificar tuberculose em imagens de radiografias de tórax.

No geral, os resultados encontrados são satisfatórios por serem próximos aos de outros trabalhos e não usarem CNN. Apesar disso, surge a oportunidade de melhorá-los, trabalhando a etapa de seleção de características com outros métodos que não necessitem de redimensionamento, por exemplo, aumentando a base usada, visto que 1600 ainda é um número baixo para atingir resultados relevantemente satisfatórios em problemas da área de saúde.

Para trabalhos futuros, uma abordagem interessante de se trabalhar é relativa a criação de um ensemble de especialistas com curvas principais, assim para cada classe haveria uma curva representativa e a classificação daria por meio de votação. Outra possibilidade seria criar uma curva principal com as imagens de Não-TB, usando a mesma ideia apresentada neste trabalho. Por fim, pode-se associar as abordagens com as curvas principais de TB e Não-TB.

REFERÊNCIAS

[1] World Health Organization (WHO), "Global Tuberculosis Reports," Disponível: https://www.who.int/teams/global-tuberculosis-programme/tb-reports

[2] World Health Organization (WHO), "Tuberculosis," Disponível: https://www.who.int/health-topics/tuberculosis#tab=tab_1

[3] Ministério da Saúde, "Manual de Recomendações e Controle da Tuberculose no Brasil 2ª ed," GOV.BR. Disponível: https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/svsa/tuberculose/manual-de-recomendacoes-e-controle-da-ttuberculose-no-brasil-2a-ed.pdf/view

[4] World Health Organization (WHO), "Latent tuberculosis infection: updated and consolidated guidelines for programmatic management," Disponível: https://www.who.int/publications/i/item/9789241550239

[5] World Health Organization, “WHO consolidated guidelines on tuberculosis: module 2: screening: systematic screening for tuberculosis disease,” Disponível: https://apps.who.int/iris/bitstream/handle/10665/340255/9789240022676-eng.pdf. [Acessado em: 14-Nov-2023].

[6] M. E. Karar, E. E.-D. Hemdan e M. A. Shouman, “Cascaded deep learning classifiers for computer-aided diagnosis of COVID-19 and pneumonia diseases in X-ray scans,” Complex & Intell. Syst., setembro de 2020. Disponível: https://doi.org/10.1007/s40747-020-00199-4

[7] T. A. Alvarenga, L. O. Santos, D. Z. Rodriguez, D. Ferreira, B. H. G. Barbosa e J. Seixas, “Unsupervised Class-Expert Learning for Supporting Covid-19 Triage Based on Computed Tomography Data,” Learn. Nonlinear Models, vol. 20, n.º 2, pp. 74–88, dezembro de 2022. Disponível: https://doi.org/10.21528/lnlm-vol20-no2-art6

[8] M. A. Ponti, G. B. P, junho de 2018. Como funciona o deep learning. arXiv preprint arXiv:1806.07908. Disponível: https://doi.org/10.48550/arXiv.1806.07908

[9] D. H. H. de Castro, D. D. Ferreira, Rodriguez Demóstenes Zegarra, “Utilização de Curvas Principais na triagem de pacientes com tuberculose,” CBIC 2023

[10] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani e A. Hooman, “An Overview of Principal Component Analysis”, J. Signal Inf. Process., vol. 04, n.º 03, pp. 173–175, 2013. Consult. 2023-11-21. Disponível: https://doi.org/10.4236/jsip.2013.43b031

[11] T. Hastie e W. Stuetzle, “Principal Curves”, J. Amer. Statistical Assoc., vol. 84, n.º 406, pp. 502–516, junho de 1989. Consult. 2023-11-21. Disponível: https://doi.org/10.1080/01621459.1989.10478797

[12] E. C. C. Moraes, D. D. Ferreira, G. B. Vitor e B. H. G. Barbosa, “Data clustering based on principal curves”, Advances Data Anal. Classification, vol. 14, n.º 1, pp. 77–96, junho de 2019. Consult. 2023-11-21. Disponível: https://doi.org/10.1007/s11634-019-00363-w

[13] J. J. Verbeek, N. Vlassis, B. Kröse, “A soft k-segments algorithm for principal curves,” Artificial Neural Networks—ICANN 2001: International Conference Vienna, Austria, August 21–25, 2001 Proceedings 11, p. 450-456, 2001.

[14] F. E. de Melo Borges, O. F. Mota, D. D. Ferreira e B. H. G. Barbosa, “One-class classifier based on principal curves”, Neural Comput. Appl., junho de 2023. Consult. 2023-11-21. Disponível: https://doi.org/10.1007/s00521-023-08721-8.

[15] F. E. M. Borges, D. D. Ferreira e J. M. Seixas, “Aprendizagem em grandes volumes de dados: Seleção de Dados para Treinamento de Máquina em Ambientes com Alta Taxa de Eventos,” in Congr. Bras. Autom. - 2020. sba, 2020. Disponível: https://doi.org/10.48011/asba.v2i1.1359

[16] H. Wang e T. C. M. Lee, “Automatic parameter selection for a k-segments algorithm for computing principal curves”, Pattern Recognit. Lett., vol. 27, n.º 10, pp. 1142–1150, julho de 2006. Consult. 2023-11-21. Disponível: https://doi.org/10.1016/j.patrec.2005.12.005

[17] D. Chicco, Siamese Neural Networks: An Overview. In: Cartwright, H. (eds) Artificial Neural Networks. Methods in Molecular Biology, vol 2190. Humana. https://doi.org/10.1007/978-1-0716-0826-5_3

[18] Lifelong Machine Learning, “Synthesis Lectures on Artificial Intelligence and Machine Learning.”

[19] XGBoost Documentation. ”XGBoost Parameters”. XGBoost. [Online]. Available: https://xgboost.readthedocs.io/en/stable/parameter.html

[20] ScienceDirect. ”Additive Model”. [Online]. Available: https://www.sciencedirect.com/topics/engineering/additive-model

[21] S. Touzani, J. Granderson e S. Fernandes, “Gradient boosting machine for modeling the energy consumption of commercial buildings”, Energy Build., vol. 158, pp. 1533–1543, janeiro de 2018. Disponível: https://doi.org/10.1016/j.enbuild.2017.11.039

[22] Y. Liu et al., “Rethinking computer-aided tuberculosis diagnosis,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, p. 2646-2655.

[23] J. Shin, S. Yoon, Y. Kim, T. Kim, B. Go e Y. Cha, “Effects of class imbalance on resampling and ensemble learning for improved prediction of cyanobacteria blooms”, Ecolog. Inform., vol. 61, p. 101202, março de 2021. Consult. 2023-11-21. Disponível: https://doi.org/10.1016/j.ecoinf.2020.101202

CONCLUSÃO

Este trabalho teve como objetivo analisar a viabilidade de usar curvas principais, conceitos de redes neurais siamesas combinadas com o XGBoost, para classificar a tuberculose em imagens de raios-X do tórax. A motivação para essa proposta foi a necessidade de uma solução que reconhecesse bem a tuberculose, uma doença grave e complexa de ser detectada, mas que também fosse leve o suficiente para ser usada em um celular, especialmente em países em desenvolvimento e de grande extensão territorial como o Brasil, onde os recursos e o acesso a técnicas mais avançadas são limitados dependendo da região. Os resultados obtidos mostraram que os algoritmos propostos combinam simplicidade e eficácia, se aproximando dos de outros trabalhos que usam a CNN. O modelo proposto pode ser útil para auxiliar os profissionais de saúde na triagem e na detecção precoce da TB, reduzindo o risco de uma triagem inadequada. No entanto, o modelo ainda tem limitações, como a dependência de mais dados, mais hiperparâmetros ou outras técnicas de pré-processamento e extração de características. Como trabalhos futuros, sugere-se testar o modelo em outras bases de dados, compará-lo com outros algoritmos, e avaliar o seu impacto na prática clínica.

REFERÊNCIAS BIBLIOGRÁFICAS

WORLD HEALTH ORGANIZATION (WHO). “Tuberculosis.”. 2019 [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tuberculosis.

MINISTÉRIO DA SAÚDE. “Manual de Recomendações para o Controle da Tuberculose no Brasil.” . (2019). [Online]. Available: https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/svsa/tuberculose/manual-de-recomendacoes-e-controle-da-ttuberculose-no-brasil-2a-ed.pdf/view.

WORLD HEALTH ORGANIZATION, “Latent tuberculosis infection: updated and consolidated guidelines for programmatic management,” Geneva, 2018, CC BY-NC-SA 3.0 IGO.

BARBOZA, M. F. X., de Souza Sampaio, V., & Endo, P. T. (2021, June). Análise e predição de incidência de casos de malária no tempo e no espaço utilizando modelos deep learning. In Anais Estendidos do XVII Simpósio Brasileiro de Sistemas de Informação (pp. 73-75). SBC

RESENDE, I. C., Cardoso, L. A., Ferreira, A. C. B., Barbosa, B. H., & Ferreira, D. D. Identificação de Grau de Risco de Pé Diabético por meio de Técnicas de Aprendizado de Máquinas. (2020). Anais da Sociedade Brasileira de Automática, 2(1).

WORLD HEALTH ORGANIZATION, “WHO consolidated guidelines on tuberculosis: module 2: screening: systematic screening for tuberculosis disease,” [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/340255/9789240022676-eng.pdf. [Accessed: 14-Nov-2023]

PONTI, Moacir Antonelli; DA COSTA, Gabriel B. Paranhos. Como funciona o deep learning. arXiv preprint arXiv:1806.07908, 2018.

HASTIE, Trevor; STUETZLE, Werner. Principal curves. Journal of the American Statistical Association, v. 84, n. 406, p. 502-516, 1989