Utilização de Curvas Principais na triagem de pacientes com tuberculose

Davi Horner Hoe de Castro
Departamento de Automática
Universidade Federal de Lavras
Lavras, Brazil
davi.castro1@estuadante.ufla.br

Danton Diego Ferreira
Departamento de Automática
Universidade Federal de Lavras
Lavras, Brazil
danton@ufla.br

Demóstenes Zegarra Rodriguez
Departamento de Automática
Universidade Federal de Lavras
Lavras, Brazil
demostenes.zegarra@ufla.br

Wilian Soares Lacerda
Departamento de Automática
Universidade Federal de Lavras
Lavras, Brazil
lacerda@ufla.br

Fernando Elias de Melo Borges
Departamento de Automática
Universidade Federal de Lavras
Lavras, Brazil
fernandoelias.mb@gmail.com

Cecilia Aparecida Santos Silva
Departamento de Automática
Universidade Federal de Lavras
Lavras, Brazil
cecilia.silva1@estudante.ufla.br

***Resumo*—A pandemia de COVID-19 mostrou que a velocidade e precisão no diagnóstico é essencial para um bom tratamento médico. Para acelerar o diagnóstico, o setor médico está se modernizando cada vez mais de forma a procurar soluções automatizadas, incrementando o uso de inteligência artificial. Uma doença extremamente negligenciada é a Tuberculose que, se não for diagnosticada no início, ela pode ser fatal ou pode causar sequelas graves que podem perdurar por muitos anos. Este trabalho visa demonstrar um método para identificar pacientes com Tuberculose (TB) por meio de Curvas Principais (CP). Para isso, utilizou-se um banco de dados disponível na literatura dividido nas seguintes classes: Pacientes com TB, pacientes com outras doenças e pacientes saudáveis. No pré-processamento foi utilizada uma Rede Neural Convolucional (CNN) já treinada, visando usar o conhecimento prévio adquirido por transferência de aprendizado, com uma *Transfer Learning* (TL), para extração de características e consequentemente facilitar a classificação. Essas características serão então analisadas usando uma validação cruzada *K-fold*, e o algoritmo k-segmentos, para criar e treinar Curvas Principais que, então, serão utilizadas na classificação das imagens. Os resultados obtidos mostraram potencial para o uso do método em futuras predições automatizadas, atingindo índices de desempenho próximos a 0,90, (90%) de acurácia.**

***Palavras-chave*—Tuberculose; Inteligência Artificial; Curva Principal; K-Segmentos, Análise de Imagens**

I. INTRODUÇÃO

A pandemia da COVID-19 demonstrou para o mundo que, mesmo com a evolução da tecnologia e aumento da automação nas mais diversas áreas, essa integração ainda tem muito espaço para evoluir. Um exemplo pertinente é o sistema da IBM Watson [1], que tem como finalidade ajudar vários tipos de empresas a automatizar seus processos, implementar soluções de Inteligência Artificial (IA), auxiliar na predição de resultados, entre outras funcionalidades. Uma dessas funcionalidades é a ferramenta Watson Health [2], que foi desenvolvida com foco no setor médico, objetivando, por exemplo, ajudar no reconhecimento de padrões de efeitos colaterais e de como um tipo de droga interage com outra, indicar formas de tratamento para doenças, criar diagnósticos por meio da análise de imagens, etc.

É importante lembrar, no entanto, que o surgimento da COVID-19 não implica no esquecimento de outras doenças já conhecidas. Segundo o relatório da OMS de 2022 [3] estima-se que 10 milhões de pessoas tenham contraído tuberculose (TB) em 2020, mas apenas 5,8 milhões foram diagnosticadas, demonstrando uma diminuição de 18% na taxa de diagnósticos em relação ao ano de 2019 devido aos *lockdowns*. O relatório ainda mostra que, pela primeira vez desde 2005 houve um aumento de mortes anuais por TB. Durante a pandemia, estima-se uma redução de 15% no número de pessoas sendo tratadas contra a variante de TB resistente às drogas, e uma diminuição de 21% nas pessoas recebendo tratamento preventivo contra TB resistente às drogas. Ao mesmo tempo, houve um corte significativo nos gastos com o foco em combater a Tuberculose [4].

Vale ressaltar que, caso a TB não for tratada nos estágios iniciais, esta representa um alto risco de mortalidade. Sendo assim, é imprescindível que o tratamento precoce seja realizado, entretanto é necessário que o diagnóstico seja feito o mais cedo possível. Um estudo em hospitais mostrou que, devido ao fato de alguns detalhes nas imagens de raio-X serem imperceptíveis ao olho humano, os radiologistas têm uma acurácia de 68,7% em relação ao exame de referência [5]. O exame de referência se refere ao melhor exame de diagnóstico disponível em condições razoáveis, mas isso no caso da TB carrega algumas complicações, pois o exame de referência no caso da TB além de extremamente caro, leva muito tempo,

pois é necessário gerar e examinar a cultura da bactéria, demandando um elevado laboratório de biossegurança. Isso demonstra porque o avanço da IA e computação para fins de diagnóstico médico está se tornando cada vez mais conhecido e relevante para a área médica.

Apesar da longa história da TB e sua prioridade no âmbito médico, o acesso a bancos de imagens de raio-X é extremamente dificultado, devido tanto à política de privacidade que os rege, quanto ao fato de que muitos são caros ou privados. Por estas razões, ainda não existe um grande número de banco de dados públicos com imagens de raio-X disponíveis para que se possa testar novas ferramentas de diagnóstico utilizando IA. Além do baixo número de repositórios públicos, a maior parte tem sua qualidade comprometida, o que pode acarretar em erros no treinamento dos modelos de aprendizagem.

Segundo o estudo reportado em [6], a pandemia de COVID-19 causou um aumento na automatização dos serviços, sobretudo, no âmbito médico. A revista *HealthTech Magazine* [7] aponta que o uso de IA na medicina está sendo imprescindível no combate à COVID-19, e um dos exemplos mais recentes disso é o algoritmo Cimatec-XCOV19 [8]. Este algoritmo foi desenvolvido para ajudar no diagnóstico da doença por meio da análise de imagens de Raio-X quando a realização do exame RT-PCR não é possível.

Tais avanços, especificamente na área de diagnóstico, se devem, principalmente, aos avanços nas técnicas de processamento de imagens nos últimos anos. A causa dessa popularização foi o crescente número de pesquisas em Redes Neurais, se destacando entre elas, as Redes Neurais Convolucionais (CNN, do inglês*Convolutional Neural Network*), onde se pode observar vários artigos que informam os seus pontos positivos quando usados na análise de imagens [9]. Sua principal vantagem é o reconhecimento das características mais importantes das imagens sem supervisão humana, o que faz com que tenha uma alta acurácia para análise das mesmas [10]. No entanto, o uso dessa técnica exige um alto poder computacional, além de demandar uma grande quantidade de dados e levar um longo tempo para treinar e validar imagens. Por precisar de um *hardware* robusto com grande capacidade em poder de Processadores Gráficos (GPUs), os sistemas embarcados devem atender a esses requisitos antes que se possa começar a explorar a eficiência de uso dessa rede.

Destarte, neste trabalho propõe-se um método de classificação de imagens de pulmão, por meio de Curvas Principais (CP) [11], [12]. As curvas principais já demonstraram, previamente, bons resultados em problemas de reconhecimento de padrões, como por exemplo na classificação de embarcações [13], na área médica foi usado na detecção e contorno de pulmões [14], entre outras pesquisas como por exemplo [15], onde se analisou a utilização do algoritmo de k-segmentos suave para achar a curva principal em diferentes tipos de banco de dados. A intenção do uso de curvas principais neste trabalho se dá, principalmente, à sua boa capacidade de representação de dados de alta dimensão e do seu baixo custo computacional em fase operacional.

O trabalho está dividido nas seguintes seções: Na seção 2, uma revisão bibliográfica da Tuberculose e os impactos que a pandemia do COVID-19 causou, como é montado uma Curva Principal usando o algoritmo de k-segmentos, e como a CP é usada para classificar uma imagem. Na seção 3, será apresentado o banco de dados que foi usado, e que pré-processamento foi realizado nos dados, e o funcionamento do projeto. Na seção 4, os resultados obtidos.

II. REVISÃO BIBLIOGRÁFICA

Esta seção visa introduzir os conceitos tratados neste trabalho. Primeiramente, delineia-se a TB, em seguida, a teoria de Curvas Principais e do algoritmo K-segmentos, e, por fim, a classificação de imagens com o uso de curvas.

A. Tuberculose

A tuberculose (TB) é uma doença transmitida pela bactéria *Mycobacterium tuberculosis*. De acordo com [16], se trata de um problema mundial que assola a humanidade há mais de 4.000 anos. Por ser transmitida pelo ar, a TB normalmente ataca o pulmão, mas pode afligir, também, por exemplo, o cérebro, o intestino, os rins e a coluna. Devido a sua alta taxa de mortalidade, em 1993 se tornou a primeira doença infecciosa a ser reconhecida como uma emergência global pela Organização Mundial de Saúde (OMS). Se não for tratada nos estágios iniciais, a TB pode levar à morte ou deixar sequelas graves nos sobreviventes. Por isto, e com a criação e adoção do tratamento precoce padronizado na década de 80, uma significativa diminuição nos casos, principalmente em países desenvolvidos, foi detectada. No entanto, em países subdesenvolvidos e em desenvolvimento, a diminuição da incidência da TB ainda tem sido lenta devido a vários fatores, que abarcam desde o clima e a infraestrutura até os aspectos socioculturais desses lugares. Estes são um dos motivos apresentados por [17] em relação à dificuldade de combater a TB no Brasil, por exemplo, continuando a ser a causa de milhões de mortes todos os anos no mundo inteiro.

A OMS indica que, para extinguir a TB até 2030, seria necessário um investimento anual de 2 bilhões de dólares. O valor máximo que foi arrecadado, no entanto, foi de 0,9 bilhão de dólares em 2020, um valor muito abaixo do que se esperava. Essa informação é discrepante se comparado aos valores disponibilizados pela plataforma de financiamento Devex [18], que informa que foi direcionado ao combate da COVID-19 o valor de 21,7 trilhões de dólares.

No ano de 2020 durante a pandemia do COVID-19, como foi reportado em [19], 1,8 milhão de pessoas morreram por COVID-19, e 1,5 milhão de pessoas morreram de tuberculose. Analisando esses números, é possível observar que houve em 2020 um número de mortes semelhantes entre Tuberculose e COVID-19, e a TB segue negligenciada comparada com a resposta que a COVID-19 causou no mundo.

Como apresentado no artigo [19], a discrepância nos investimentos ao combate à TB e à COVID-19, não esta ligado ao número de mortos, mas sim a que populações são mais afetadas por essas doenças, onde a COVID-19 é difundida

igualmente em todos os países, desenvolvidos e em desenvolvimento, a TB afeta primeiramente os países mais pobres e suas regiões mais vulneráveis.

O fator mais pertinente na análise dessa diferença é a demonstração de que a TB ainda é uma doença extremamente negligenciada, mesmo sendo considerada letal. E, portanto, é imprescindível que pesquisas busquem novas formas de tratamento, medicação, vacinas e diagnósticos sejam realizadas com urgência para que se possa salvar o máximo de vidas possível.

B. Curva Principais - K-Segmentos

A técnica de Curvas Principais foi definida inicialmente por [11] como curvas unidimensionais que atravessam o meio de um conjunto de dados em um espaço multidimensional, fazendo uma representação compacta do mesmo. Usando tal proposta como base, algoritmos alternativos foram explorados e desenvolvidos a fim de se obter melhores extrações das Curvas Principais, que tenham, ao mesmo tempo, uma convergência prática e um desempenho computacional satisfatório. Há, por exemplo, o algoritmo usado neste trabalho, conhecido como algoritmo K-segmentos [12]. Este algoritmo possui menor influência a mínimos locais e tem convergência prática garantida, fornecendo, assim, robustez ao método.

Seu funcionamento se divide nos passos representado por esse fluxograma na figura 1 que foi extraído de [20]:

(0) Inicialmente, precisa-se obter o primeiro segmento utilizando todo o conjunto de dados do sistema. O segmento é obtido na direção da primeira componente principal com comprimento de 3/2 de desvio padrão dos dados.

(1) Um segundo segmento é adicionado a um novo agrupamento feito por meio do algoritmo k-means, com base nas regiões de Voronoi. Neste contexto, essas regiões são onde os eventos de um dado agrupamento estão em maior proximidade do centro regional do que dos segmentos da curva. Depois da obtenção do segundo segmento, o cálculo do primeiro é novamente realizado em razão de uma modificação no seu agrupamento. Esse mesmo processo é feito para os demais segmentos, adicionando-se um novo segmento e realizando o recálculo dos segmentos onde houve mudança.

(2) O teste de convergência do algoritmo é realizado de dois modos. Primeiro, verifica-se se o número de segmentos k alcançou o valor máximo de segmentos esperado pelo usuário *(K_{max})*, ou se a região de Voronoi possui menos de 3 amostras. Se estiver em desacordo com as duas condições estabelecidas, o algoritmo retorna ao Passo 1.

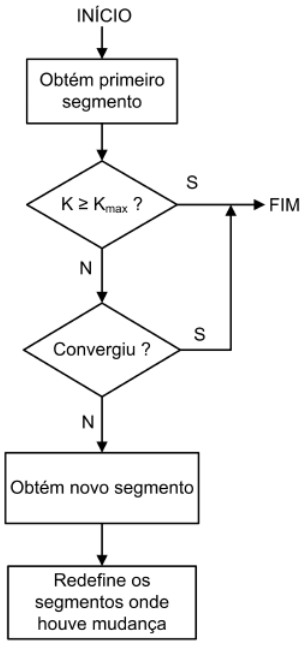

Fig. 1. Fluxograma do algoritmo k-segmentos para obtenção da Curva Principal. Fonte: Borges et al. (2020) [20]

III. Método Proposto

Nesta seção, o banco de dados, seu pré-processamento e o funcionamento do algoritmo são apresentados.

A. Banco de dados

O banco de dados utilizado pela presente pesquisa foi o TBX11K [5], ele totaliza 11200 imagens. Tal número indica uma quantidade razoável para o treinamento de modelos de inteligência artificial. Outros aspectos positivos desse banco dados que impactou na nossa escolha foi que o mesmo é altamente padronizada, e permite treinamento de detectores sofisticados e apresenta uma anotação de caixa delimitadora que deixa as imagens do mesmo tamanho melhorando assim o treinamento de Redes Neurais Convolucionais, tendo sido projetado especificamente para ajudar no diagnóstico de TB, aumentando, assim, a precisão do treinamento.

Esse banco de dados é dividido em várias classes, de acordo com a situação do paciente e a severidade da TB. Contudo, de maneira a tornar o método mais objetivo, foi feita uma generalização do banco de dados, reduzindo o número de classes para 3, sendo: 1200 imagens de pacientes com TB, 5000 imagens de pacientes com outras doenças e 5000 imagens de pacientes saudáveis. Observando o número de amostras por classe, observa-se um significativo desbalanceamento na base dados.

Um desbalanceamento nos dados pode gerar uma série de problemas, como por exemplo, não conseguir detectar uma doença. Dessa forma é necessário adaptar o banco de dados para diminuir o desbalanceamento, para isso foi identificado a classe com o menor número de imagens e com a maior relevância para detectar TB ativo, desta forma foi escolhido a classe com TB ativo, que esta possui apenas 924 imagens. Com isso definimos um limite para as classes de 800 imagens cada. Dessa forma depois do balanceamento cada classe ficará com apenas 800 imagens para ser usado no projeto.

B. Pré-processamento com Transfer Learning

Antes de ser utilizado no projeto, o banco de dados balanceado passa por um pré-processamento que consiste no uso de uma CNN pré-treinada, com o *Transfer Learning* (TL),

para que seja possível capturar as características das imagens. Como demonstrado pelos trabalhos realizados em [21] e pelo [22], é possível usar uma TL em um modelo já existente para conseguir resultados melhores, tanto em acurácia quanto em eficiência, caso a limitação de um banco de dados pequeno esteja presente. No contexto deste trabalho, a limitação está mais relacionada com a estrutura de *hardware* para treinar uma rede CNN para o conjunto de dados utilizado e, portanto, foi realizado o uso de uma TL.

A CNN utilizada foi a Resnet-18. Ela foi escolhida pelos seguintes motivos: 1º na pesquisa do banco [5] ela já estava sendo usada, para o mesmo propósito, então foi possível aproveitar essa etapa em mais de um processo; 2º pois esse tipo já é bastante estudado e tem comprovadamente bons resultados [23]; e 3º porque as camadas intermediárias tem 512 neurônios que é um número relativamente pequeno se comparado à outras CNN.

Para reduzir a dimensionalidade das características extraídas pelas CNN, foi aplicada uma Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) com variância de 0,95.

C. Curvas Principais como Classificador

Para o projeto do classificador, foi utilizada uma abordagem supervisionada por meio de Curvas Principais. Para isso, uma Curva Principal foi criada para cada uma das classes referentes às condições dos pacientes.

Depois da criação das CP, para cada nova imagem a ser analisada, será calculada a distância entre a imagem e cada uma das curvas. A classificação das imagens é a mesma da curva mais próxima.

Para ilustrar esse processo, é possível utilizar a Figura 2, onde "$_f$", "$_i$" e "$_h$" representam curvas principais em um espaço de features e analisando a imagem "$_K$", é possível calcular a distância "$_{Kf}$" que representa a distância entre "$_K$" e a curva "$_f$"; "$_{Ki}$" que representa a distância entre "$_K$" e a curva "$_i$"; e "$_{Kh}$" que representa a distância entre "$_K$" e a curva "$_h$". Descobrindo essas distâncias é possível analisar e classificar "$_K$" como pertencente a classe "$_h$".

D. Projeto do Classificador

As características depois do pré-processamento são representadas por um vetor de características onde, cada característica representa uma componente principal. Essa lista será separada em dois grupos, na proporção (80%-20%). O grupo de 20% será usado para validar o classificador, e o grupo com 80% será usado para o treinamento de uma validação cruzada do tipo *K-Fold* com divisão estratificada de tamanho 5. O *Strat K-Fold* é um método de fazer a validação cruzada dividindo os dados em K grupos, onde K tem que ser no mínimo 2. O lado positivo dessa validação é que mantém a porcentagem de dados entre as classes.

Cada um dos *Folds* será treinado usando o método demonstrado por [12], que consiste em construir Curvas Principais usando segmentos de reta que represente cada classe. Em seguida, com estas curvas elas serão validadas testando a sua

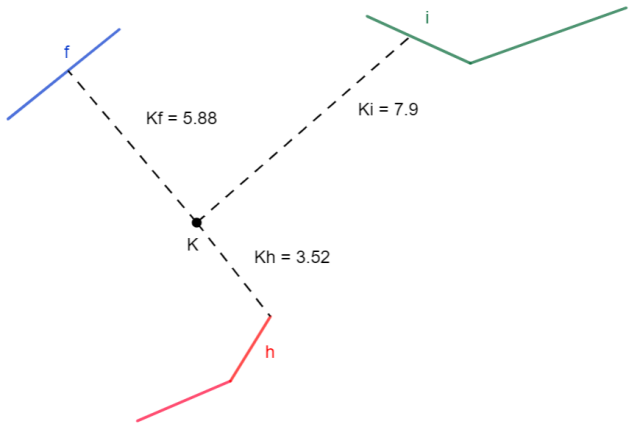

Fig. 2. Exemplo do uso de Curvas Principais como classificador.

predição, utilizando-se ainda dos dados no grupo de 80%. Com os resultados obtidos será montado uma matriz confusão para cada *Fold*, descobrindo assim a acurácia de cada *Fold*. Também será calculado os valores de *Precision* e *Recall* para cada classe. Analisando as matrizes confusões, será feito uma média das matrizes para poder analisar a consistência dos resultados.

Para certificar que o treinamento obteve predições satisfatórias, serão selecionadas as CP que obtiverem a melhor acurácia no treinamento. A partir desse ponto as CP escolhidas serão testadas novamente agora usando o grupo de dados com 20% e novamente será montado uma matriz confusão, descobrindo assim uma nova acurácia e também será calculado os valores de *Precision* e *Recall* para cada classe.

Esse projeto foi criado utilizando a linguagem de programação *Python* e a IDE *Spyder*, e testado em um computador com 16 GB de RAM e um processador i7-12700H 2.30 GHz.

IV. Resultados e Discussão

Para entender a métrica utilizada é preciso primeiro entender a matriz confusão. A matriz de confusão, tem a função de permitir visualizar o desempenho de um algoritmo. Cada linha da matriz representa as instâncias em uma classe real enquanto cada coluna representa as instâncias em uma classe prevista. Na Tabela I, é mostrado o funcionamento de uma matriz de confusão.

A matriz confusão nesse projeto está sendo usada na configuração de uma análise preditiva. Neste formato a matriz tem a seguinte configuração, duas linhas e duas colunas que informam o número de verdadeiros positivos, falsos negativos, falsos positivos e verdadeiros negativos. Isso permite uma análise mais detalhada que a observação de classificações corretas (Acurácia). A Acurácia gera dados enganosos quando o banco de dados é desbalanceado, o que reforça o balanceamento feito anteriormente.

As métricas utilizadas para analisar a classificação são a Acurácia (do inglês *Accuracy*), Precisão (do inglês *Precision*)

Tabela I
MATRIZ CONFUSÃO.

		Valor Real	
		Negativo	**Positivo**
Valor Previsto	**Negativo**	Verdadeiro Negativo	Falso Negativo
	Positivo	Falso Positivo	Verdadeiro Positivo

e a Revocação (do inglês *Recall*). Estas podem ser descobertas usando , respectivamente, as seguintes equações 1, 2, 3.

A acurácia é o número de pontos de dados previstos corretamente de todos os pontos de dados. Sua formulação é descrita pela equação 1.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precisão é definida pelo número de verdadeiros positivos levando em conta todos os documentos recuperados. A precisão e a revocação são usadas com frequência juntas pois elas são complementares. Conforme observado na equação 2.

$$P = \frac{TP}{TP + FP} \tag{2}$$

É imprescindível ficar atento que o significado e o uso de precisão na área de recuperação da informação é diferente da definição de acurácia e precisão dentro de outros ramos da ciência e tecnologia.

Em uma classificação binária a revocação também pode ser chamada de sensibilidade. Onde ela pode ser entendida como a probabilidade de que um documento relevante seja obtido pela consulta. A formulação da revocação é descrita na equação 3.

$$R = \frac{TP}{TP + FN} \tag{3}$$

Na área médica é extremamente importante analisar a revocação, pois ela indica a porcentagem de pacientes com uma doença que recebeu um diagnóstico negativo, podendo causar a morte desse paciente por falta de tratamento.

É frequente alcançar uma revocação de 100% ao retornar todos os documentos em resposta a uma consulta. Mostrando assim que a revocação por si só não é suficiente, precisando assim da necessidade de medir também o número de documentos não relevantes, por exemplo, calculando a precisão.

A Tabela II mostra os valores médios e desvios-padrão dos *folds* encontrados, enquanto que a Tabela III mostra os resultados para a melhor curva principal, ou seja, aquela que alcançou as melhores métricas. Nota-se que uma acurácia de quase 90% foi alcançada. O *Precision* e *Recall* mostram altos valores para as classes TB e doente, o que indica que eles classificam corretamente os dados. Ao verificar o valor mais baixo na precisão para a classe saudável no conjunto de teste, nota-se que o modelo a prevê bem, mas com menor especificidade. Os baixos valores no desvio padrão indicam que esses resultados são consistentes e que não há uma grande variação de um *fold* para o outro. Os resultados da validação foram similares aos resultados de teste, indicando boa generalização.

Tabela II
RESULTADOS DA ACURÁCIA, PRECISÃO E REVOCAÇÃO DA MÉDIA E DESVIO PADRÃO DE TODAS AS CURVAS PRINCIPAIS.

Conjunto	**Medida**	**K-seg**
Teste	Accuracy	0.88 ± 0.01
	Precision tb	0.90 ± 0.02
	Recall tb	0.84 ± 0.03
	Precision doente	0.92 ± 0.02
	Recall doente	0.89 ± 0.01
	Precision saudável	0.77 ± 0.03
	Recall saudável	0.97 ± 0.01

Tabela III
RESULTADOS DA ACURÁCIA, PRECISÃO E REVOCAÇÃO DA VALIDAÇÃO DA MELHOR CURVA PRINCIPAL.

Conjunto	**Medida**	**K-seg**
Validação	ACC	0.89
	Precision tb	0.88
	Recall tb	0.82
	Precision doente	0.90
	Recall doente	0.90
	Precision saudável	0.89
	Recall saudável	0.95

As imagens das matrizes de confusão foram geradas usando a biblioteca *Seaborn* [24] para melhor visualização. Observando as figuras 3 e 4, observa-se que o método proposto apresentou bom resultado de precisão tanto para validação quanto para teste, para as três classes consideradas. Ainda analisando as matrizes podemos perceber que os erros estão espalhados entre as classes indicando que o balanceamento ajudou a impedir que os erros sejam tendenciosos para uma classe em específico. Também é possível observar que dada a restrição na quantidade de imagens no banco de dados para a classe TB, as matrizes apontam que a classe TB é a classe que possui sua classificação mais depreciada, onde imagens de TB são confundidas com imagens de pulmões sadios e doentes.

Por causa de que, a classificação faz apenas o cálculo da distância da amostra para a curva, e não a montagem das curvas, gera a simplificação do programa deixando ele com uma complexidade menor e com um desempenho melhor.

Fazendo uma comparação com os resultados obtidos no artigo [25], que é revisão recente sobre métodos computacionais de apoio ao diagnóstico da TB, voltado para trabalhos que utilizaram o mesmo banco de dados da presente proposta, percebe-se que os resultados obtidos de acurácia (0,89), precisão média (0,90) e Recall média (0,84), apesar de serem inferiores aos resultados do método proposto em [25] (acurácia igual a 0,93, precisão média igual a 0,92 e Recall média igual a 0,92), são bastante competitivos. Investigando os demais métodos usados para comparação em [25], nota-se

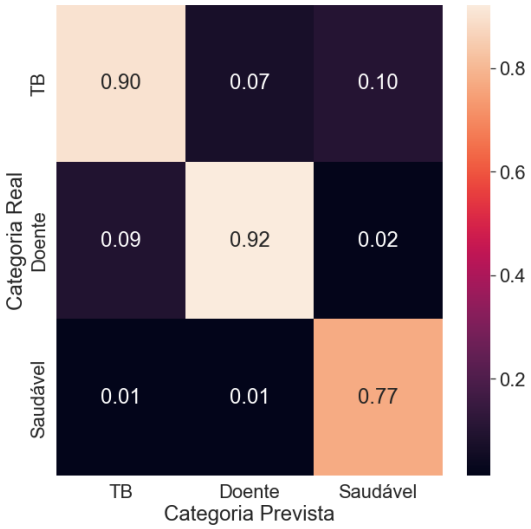

Fig. 3. Matriz de Confusão dos valores de precisão da média das CPs normalizada.

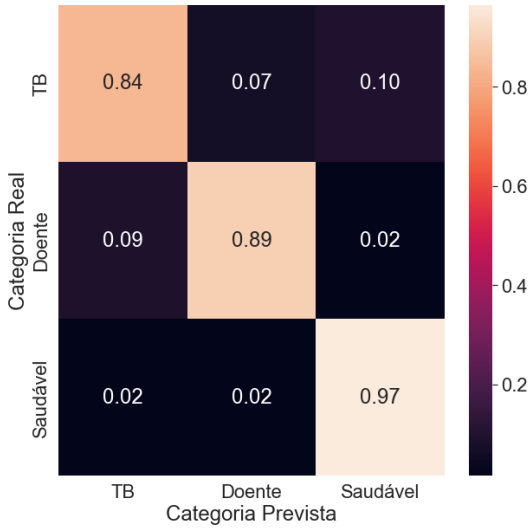

Fig. 5. Matriz de Confusão dos valores de revocação da média das CPs normalizada.

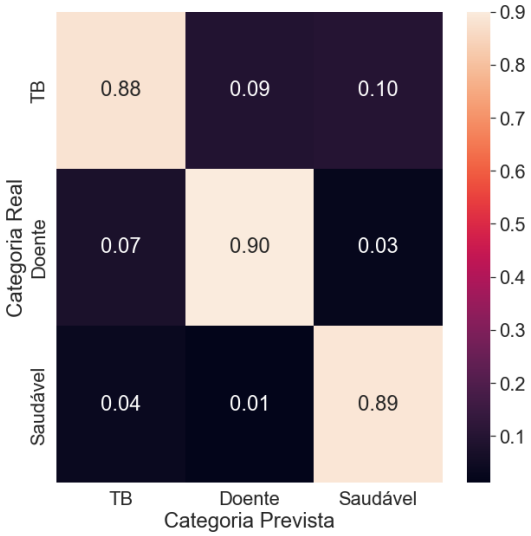

Fig. 4. Matriz de Confusão dos valores de precisão da melhor CP normalizada.

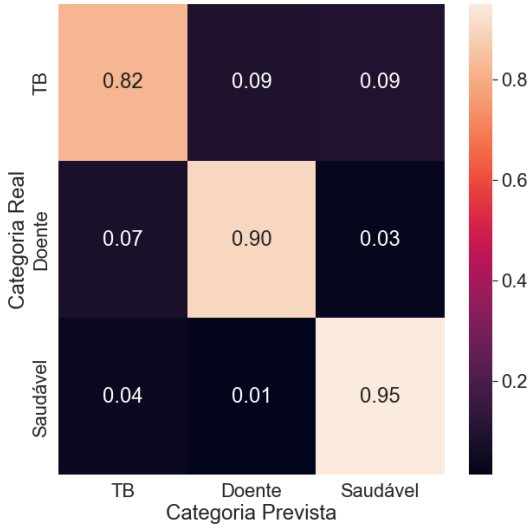

Fig. 6. Matriz de Confusão dos valores de revocação da melhor CP normalizada.

que a acurácia aqui obtida é mediana e a precisão média é maior.

V. CONCLUSÃO

O presente trabalho teve por objetivo, fazer uma análise inicial se a utilização de Curvas Principais, na triagem de pacientes tem um resultado competitivo com as tecnologias mais pesquisadas na literatura.

Os resultados no presente trabalho, comprova que é possível fazer a utilização das Curvas Principais para facilitar no diagnóstico de TB.

As dificuldades encontradas em relação aos bancos de dados, no entanto, demonstram que é necessário que uma solução seja buscada, em conjunto com hospitais e pacientes, de forma a melhorar a disponibilidade de dados para treinamentos de modelos de aprendizagem.

Outro aspecto pertinente visto ao longo desta pesquisa é que a TB, mesmo tendo sua letalidade tão conhecida, não tem sido tratada com a devida seriedade, especialmente após o surgimento da COVID-19. A pandemia, no entanto, exigiu que o mundo investisse mais na automatização da área médica, principalmente no âmbito de diagnósticos, portanto espera-se que esses avanços aos poucos sejam adaptados para o tratamento da TB e de outras doenças.

Futuramente, é possível realizar uma pesquisa prática, com o auxílio de hospitais, para averiguar se essa classificação teria os mesmos resultados se utilizado fora de um ambiente controlado, e quais os impactos reais no processo de diagnosticar

pacientes esse método traria para um hospital.

Ao analisar os resultados do método proposto com outros métodos, como por exemplo, a CNN, é necessário não apenas observar no quesito acurácia, mas também a exigência computacional do *hardware*. Conforme explicado previamente, a tecnologia predominantemente usada na área é a CNN, entretanto ainda não existem estudos conclusivos da viabilidade desse modelo em sistemas *IoT* (Internet of Things), embarcados e em aplicativos de *smartphones*. Dessa maneira o modelo utilizando Curvas Principais com uma exigência de *hardware* muito inferior às CNN possui vantagem em aplicações com restrições computacionais.

AGRADECIMENTOS

Eu gostaria de agradecer a Deus por tudo; à minha família, pelo incentivo que tive todos esses anos; aos meus amigos que me deram suporte e me ajudaram nessa jornada; ao Dr. Alessandro Wasum Mariani, que permitiu que eu estivesse aqui hoje e me inspirou a escolher este tema; e por fim à toda Universidade Federal de Lavras e aos meus orientadores, que tanto me apoiaram nessa jornada. E gostaria de agradecer também a CNPq, Fapemig e CAPES por apoiar o trabalho.

REFERÊNCIAS

[1] IBM, "Ibm watson is ai for smarter business," 2022, acesso em: 20 de Setembro de 2022. [Online]. Available: https://www.ibm.com/watson

[2] I. Watson, "O watson health é a saúde mais inteligente," 2022, acesso em: 20 de Setembro de 2022. [Online]. Available: https://www.ibm.com/br-pt/watson-health

[3] W. H. Organization, "Global tuberculosis report 2022," 2022. [Online]. Available: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2022

[4] S. S. Madhukar Pai, Tereza Kasaeva, "Covid-19's devastating effect on tuberculosis care — a path to recovery," *The New England Journal of Medicine*, 01 2022.

[5] Y. Liu, Y.-H. Wu, Y. Ban, H. Wang, and M.-M. Cheng, "Rethinking computer-aided tuberculosis diagnosis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2646–2655.

[6] G. H. Erik R. Peterson, Terence Toland, "Robots vs. covid-19: how the pandemic is accelerating automation," https://www.kearney.com/web/global-business-policy-council/article/-/insights/robots-vs-covid-19-how-the-pandemic-is-accelerating-automation, 2020.

[7] A. Steger, "Healthcare automation matters more than ever during a pandemic," https://healthtechmagazine.net/article/2020/07/healthcare-automation-matters-more-ever-during-pandemic-perfcon, 2020.

[8] A. Furtado, C. Purificação, R. Badaró, and E. G. Sperandio Nascimento, "A light deep learning algorithm for ct diagnosis of covid-19 pneumonia," *Diagnostics (Basel, Switzerland)*, vol. 12, 06 2022.

[9] S. Cantrell, "Top 3 most popular neural networks," 2018, acesso em: 20 de Setembro de 2022. [Online]. Available: https://www.excella.com/insights/top-3-most-popular-neural-networks

[10] L. Alzubaidi, J. Zhang, A. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, 03 2021.

[11] T. Hastie and W. Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478797

[12] J. Verbeek, N. Vlassis, and B. Kröse, "A k-segments algorithm for finding principal curves," *Pattern Recognition Letters*, vol. 23, no. 8, pp. 1009–1017, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865502000326

[13] H. L. Fernandez, "Classificação de navios baseada em curvas principais," Master's thesis, Engenharia de Sistemas e Automação - Universidade Federal do Rio de Janeiro, 2005.

[14] T. Peng, Y. Wang, T. Xu, L. Shi, J. Jiang, and S. Zhu, "Detection of lung contour with closed principal curve and machine learning," *Journal of Digital Imaging*, vol. 31, 02 2018.

[15] E. C. C. Moraes, "Método não supervisionado baseado em curvas principais para reconhecimento de padrões," Master's thesis, Universidade Federal de Lavras, 2015.

[16] K. Zaman, "Tuberculosis: a global health problem," *Journal of health, population, and nutrition*, vol. 28, no. 2, p. 111—113, April 2010. [Online]. Available: https://europepmc.org/articles/PMC2980871

[17] A. O. Cortez, A. C. d. Melo, L. d. O. Neves, K. A. Resende, and P. Camargos, "Tuberculosis in brazil: one country, multiple realities," *Jornal brasileiro de pneumologia : publicacao oficial da Sociedade Brasileira de Pneumologia e Tisilogia*, vol. 47, no. 2, p. e20200119, 2021. [Online]. Available: https://europepmc.org/articles/PMC8332839

[18] L. Cornish, "Interactive: Who's funding the covid-19 response and what are priorities," 2021, acesso em: 20 de Setembro de 2022. [Online]. Available: https://tinyurl.com/34ex556k

[19] E. L. Maciel, J. E. Golub, J. R. L. E. Silva, and R. E. Chaisson, "Tuberculosis: a deadly and neglected disease in the covid-19 era," *Jornal brasileiro de pneumologia : publicacao oficial da Sociedade Brasileira de Pneumologia e Tisilogia*, vol. 48, no. 3, p. e20220056, June 2022. [Online]. Available: https://doi.org/10.36416/1806-3756/e20220056

[20] L. M. Borges, F. E. M. Borges, D. A. Ribeiro, A. W. M. Pinto, and D. D. Ferreira, "Uso de curvas principais na classificação de falhas em motor de indução trifásico," *Congresso Brasileiro de Automática*, 12 2020, acesso em: 20 de Setembro de 2022.

[21] M. Hussain, J. Bird, and D. Faria, "A study on cnn transfer learning for image classification," in *A Study on CNN Transfer Learning for Image Classification*, 06 2018, p. 1.

[22] K. Wang, X. Gao, Y. Zhao, X. Li, D. Dou, and C.-Z. Xu, "Pay attention to features, transfer learn faster cnns," in *International Conference on Learning Representations*, 2020, p. 1, acesso em: 20 de Setembro de 2022. [Online]. Available: https://openreview.net/forum?id=ryxyCeHtPB

[23] S. Ayyachamy, V. Alex, M. Khened, and G. Krishnamurthi, "Medical image retrieval using Resnet-18," in *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications*, P.-H. Chen and P. R. Bak, Eds., vol. 10954, International Society for Optics and Photonics. SPIE, 2019, p. 1095410, acesso em: 20 de Setembro de 2022. [Online]. Available: https://doi.org/10.1117/12.2515588

[24] "seaborn: statistical data visualization," 2022, acesso em: 20 de Setembro de 2022. [Online]. Available: https://seaborn.pydata.org/

[25] Y. Liu, Y.-H. Wu, S.-C. Zhang, L. Liu, M. Wu, and M.-M. Cheng, "Revisiting Computer-Aided Tuberculosis Diagnosis," *arXiv e-prints*, p. arXiv:2307.02848, Jul. 2023.

[26] G. G. M. d. Rocha and J. B. d. O. e. S. Filho, "Classificação de contatos baseada em curvas principais utilizando o processador nios ii," *Congresso Brasileiro de Automática*, 09 2014.