DAVI HORNER HOE DE CASTRO

UTILIZAÇÃO DE CURVAS PRINCIPAIS NA TRIAGEM DE PACIENTES COM TUBERCULOSE

LAVRAS – MG

2022

DAVI HORNER HOE DE CASTRO

UTILIZAÇÃO DE CURVAS PRINCIPAIS NA TRIAGEM DE PACIENTES COM TUBERCULOSE

Trabalho de Conclusão de Curso apresentada à Universidade Federal de Lavras, como parte das exigências do Curso de Graduação em Ciência da Computação, para obtenção do título de Bacharel.

gov.br
Documento assinado digitalmente
DANTON DIEGO FERREIRA
Data: 23/09/2022 16:29:16-0300
Verifique em https://verificador.iti.br

Prof. DSc. Danton Diego Ferreira

Orientador

Prof. DSc. Demóstenes Zegarra Rodriguez

Coorientador

LAVRAS – MG

2022

Ficha catalográfica elaborada pela Coordenadoria de Processos Técnicos

da Biblioteca Universitária da UFLA

Castro, Davi Horner Hoe de

Utilização de Curvas Principais na triagem de pacientes com tuberculose / Davi Horner Hoe de Castro. 2ª ed. rev., atual. e ampl. – Lavras : UFLA, 2022.

26 p. : il.

TCC(Graduação)–Universidade Federal de Lavras, 2022.

Orientador: Prof. DSc. Danton Diego Ferreira.

Bibliografia.

1. Tuberculose. 2. Inteligência Artificial. 3. Curva Principal. – Normas. I. Universidade Federal de Lavras. II. Título.

CDD-808.066

DAVI HORNER HOE DE CASTRO

UTILIZAÇÃO DE CURVAS PRINCIPAIS NA TRIAGEM DE PACIENTES COM TUBERCULOSE

Trabalho de Conclusão de Curso apresentada à Universidade Federal de Lavras, como parte das exigências do Curso de Graduação em Ciência da Computação, para obtenção do título de Bacharel.

APROVADA em 16 de Setembro de 2022.

Prof. DSc. Danton Diego Ferreira	UFLA
Prof. DSc. Demóstenes Zegarra Rodriguez	UFLA
Prof. DSc. Wilian Soares Lacerda	UFLA
MSc. Fernando Elias de Melo Borges	UFLA

Prof. DSc. Danton Diego Ferreira
Orientador

Prof. DSc. Demóstenes Zegarra Rodriguez
Co-Orientador

LAVRAS – MG
2022

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.

Tudo o que temos de decidir é o que fazer com o tempo que nos é dado.
(Gandalf)

RESUMO

A pandemia 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 que está procurando soluções automatizadas e que se utilizam cada vez mais da inteligência artificial. Uma doença extremamente negligenciada é a tuberculose que apesar do fato se a tuberculose não for diagnosticada no início ela pode ser fatal e pode causar sequelas graves que podem perdurar por muitos anos. Este trabalho visa demonstrar um método para identificar pacientes com tuberculose (TB). Utilizando um banco de dados disponível na literatura dividido nas seguintes classes: Pacientes com TB, pacientes com outras doenças e pacientes saudáveis. Esse banco de dados é tratado por um pré-processamento de uma Rede Neural Convolucional (CNN) já treinada, ou seja, há a transferência de aprendizado, com *Transfer Learning* (TL), para extrair características. 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

ABSTRACT

The pandemic has shown that speed and accuracy in diagnosis is essential for good medical treatment. To speed up diagnosis, the medical sector is increasingly modernizing so that it is looking for automated solutions that are increasingly using artificial intelligence. An extremely neglected disease is tuberculosis despite the fact that if tuberculosis is not diagnosed early it can be fatal and can cause severe long-term sequelae. This work aims to demonstrate a method to identify patients with tuberculosis (TB). Using a database available in the literature divided with the following classes: With TB patients, patients with other diseases and healthy patients. This database will be treated with a pre-processing of an already trained Convolutional Neural Network (CNN), that is, there is a transfer of learning using the method of Transfer Learning (TL), to extract its characteristics. These characteristics will then be analyzed using the K-fold cross validation, and the algorithm k-segments to create and train Principal Curves, which will then be used in the classification of the images. The results obtained showed potential use of the method in future automated predictions, reaching performance indexes close to 0.90 (90%) of accuracy.

Keywords: Tuberculosis, Artificial Intelligence, Principal Curves

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 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 (IBM, 2022), 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 (WATSON, 2022), 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. 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*. Complicando mais a situação 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 a mesma ao mesmo tendo houve um corte significativo nos gastos com o foco em combater a doença (PAI TEREZA KASAEVA, 2022).

Não se pode esquecer que caso a TB não for tratada nos estados iniciais, esta representa um alto risco de mortalidade. Sendo assim, é imprescindível que o tratamento precoce seja realizado, mas para isso, 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 (LIU et al., 2020). No entanto, além de extremamente caro, o exame de referência leva tempo, pois é necessário gerar e examinar a cultura da bactéria, além de demandar um laboratório de biossegurança nível-3. Isso demonstra porque o avanço da IA e computação para fins de diagnóstico 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 essas 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 diagnósticos utilizando IA, e menos ainda bancos de dados de boa qualidade para evitar erros no treinamento.

O que muitos analistas têm percebido, como (PETERSON TERENCE TOLAND, 2020), que a pandemia está causando um aumento na automatização de empregos e serviços em muitas áreas, principalmente no âmbito médico. A revista *HealthTech Magazine* (STEGER, 2020) 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 (FURTADO et al., 2022), este algoritmo foi desenvolvido para ajudar no diagnóstico da doença através da análise de imagens de Raio-X quando a realização do exame RT-PCR não é possível.

Esses tipos de avanço, mais especificamente na área de diagnóstico, se devem principalmente ao fato de que nos últimos anos ocorreu um grande avanço na área de análise de imagens. A causa dessa popularização foi o crescente número de pesquisa em Redes Neurais, se destacando entre elas a Rede Neural Convolucional (CNN, do inglês*Convolutional Neural Network*), onde podemos observar vários artigos que informam os seus pontos positivos se usados na análise de imagens (CANTRELL, 2018). Como apontado por (ALZUBAIDI et al., 2021), sua principal vantagem é o fato de que ela consegue reconhecer as características mais importantes das imagens sem supervisão humana, o que faz com que tenha uma alta acurácia para análise das mesmas. 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 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, utilizando uma CNN combinada com uma *Transfer Learning* (TL) para melhorar a captura das características das mesmas e usá-las para montar as Curvas Principais (HASTIE; STUETZLE, 1989) por meio do algoritmo k-segmentos (VERBEEK; VLASSIS; KRöSE, 2002). As curvas principais já demonstraram, previamente, bons resultados em problemas de reconhecimento de padrões, como por exemplo na classificação de embarcações (FERNANDEZ, 2005), na área médica foi usado na detecção e contorno de pulmões (PENG et al., 2018), entre outras pesquisas como por exemplo ("ROCHA; FILHO, 2014) e (MORAES, 2015). 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.

2 REVISÃO BIBLIOGRÁFICA

Esta seção visa introduzir brevemente os conceitos tratados no trabalho. Primeiramente, delineia-se a TB, em seguida a teoria de Curvas Principais e do método K-segmentos, e, por fim, a classificação de imagens com o uso de curvas.

2.1 Tuberculose

A tuberculose (TB) é uma doença transmitida pela bactéria *Mycobacterium tuberculosis*.De acordo com (ZAMAN, 2010), 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, e por isso, 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 (CORTEZ et al., 2021) 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.

Como mostra o artigo (MACIEL et al., 2022), a diferença de investimentos no combate à TB e à COVID-19 é um aspecto surpreendente. 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 (CORNISH, 2021), que informa que foi direcionado ao combate da COVID-19 o valor de 21,7 trilhões de dólares.

Ainda não foram disponibilizados os dados de 2021, mas usando os dados disponíveis do ano de 2020 e levando em consideração as diferenças que ambas as doenças possuem, tanto em taxas de transmissão e área mais afetadas, temos os seguintes dados. Apenas no ano de 2020 durante a pandemia do COVID-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 mesmo com o financiamento em massa para combater o COVID-19.

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.

2.2 K-Segmento - Curva Principal

A técnica de Curvas Principais foi definida inicialmente por (HASTIE; STUETZLE, 1989) 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 (VERBEEK; VLASSIS; KRöSE, 2002). Este algoritmo possui menor influência a mínimos locais e tem convergência prática garantida, fornecendo, assim, robustez ao método.

Seu funcionamento se divide nos seguintes passos e por esse fluxograma extraído de ("BORGES et al., 2020):

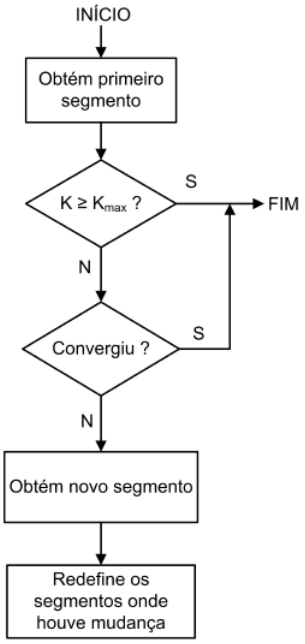
(0) Inicialmente, precisa-se obter o primeiro segmento utilizando todo o conjunto de dados do sistema. O segmento é obtido na direção da primeira componente principal com comprimento de 3/2 de desvio padrão dos dados.

(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 o maior agrupamento concebido possui, ao menos, 3 segmentos. Se estiver em desacordo com as duas condições estabelecidas, o algoritmo retorna ao Passo 1.

Figura 2.1 – Fluxograma do algoritmo k-segmentos para obtenção da Curva Principal.



Fonte: Borges et al. (2020)

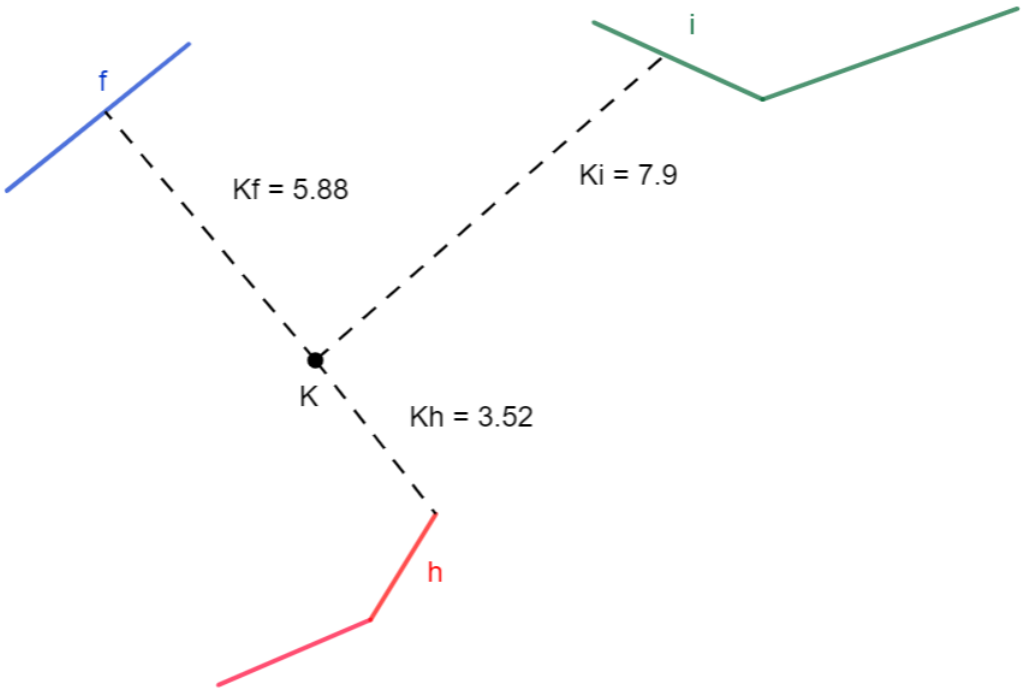
2.2.1 Curvas Principais como Classificador

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 curvas principais, 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 2.2, 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 "$_h$".

Figura 2.2 – Exemplo do uso de Curvas Principais como classificador.



Fonte: do Autor.

3 MÉTODO PROPOSTO

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

3.1 Base de dados

O banco de dados utilizado pela presente pesquisa foi o TBX11K (LIU et al., 2020), ele totaliza 11200 imagens. Tal número indica uma quantidade razoável para o treinamento de modelos de inteligência artificial. Outro aspecto positivo é que o mesmo apresenta uma anotação de caixa delimitadora que melhora 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, mas é feito uma generalização do banco de dados que deixa apenas 3 classes com as seguintes quantidades de imagens para cada classe: 1200 imagens de pacientes com TB, 5000 imagens de pacientes com outras doenças e 5000 imagens de pacientes saudáveis. Observando esses dados fica evidente que o banco de dados está profundamente desbalanceado e precisa ser adaptado, que envolve em identificar a classe com TB ativo, que esta possui apenas 924 imagens, com base nessa informação e também no ponto de que o foco do projeto é voltado para classificar pessoas com TB ativo, será definido como limite das classes 800 imagens. Dessa forma depois do balanceamento cada classe ficará com apenas 800 imagens para ser usado no projeto.

3.2 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, como *Transfer Learning* (TL), para que seja possível capturar as características das imagens. Como demonstrado pelo (HUSSAIN; BIRD; FARIA, 2018) e pelo (WANG et al., 2020), 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, o uso de uma TL preferido.

A CNN utilizada foi a Resnet-18. Ela foi escolhida pelos seguintes motivos: 1º na pesquisa do banco (LIU et al., 2020) 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 (AYYACHAMY et al., 2019), e 3º porque as camadas intermediárias tem 512 neurônios que é um número relativamente pequeno perto do que as outras CNN possuem. Foi utilizada uma Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) usando uma variância de 0,95, pois, por causa da enorme quantidade de características presentes nas imagens é necessário reduzi-las para aumentar a sua interpretabilidade ao mesmo tempo conservando informações valiosas que serão usadas na hora de montar as Curvas Principais.

3.3 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 (VERBEEK; VLASSIS; KRöSE, 2002), que consiste em construir Curvas Principais usando segmentos de reta que represente cada classe. Em seguida com essas curvas elas serão validadas testando a sua 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* e 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, será selecionado a Curvas Principais que obteve a melhor acurácia no treinamento. A partir desse ponto essa Curva Principal será testada 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.

4 RESULTADOS E DISCUSSÃO

Para entender a métrica utilizada é preciso primeiro entender a matriz confusão 4.1. 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.

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

Tabela 4.1 – Matriz Confusão

		Valor Real	
		Negativo	**Positivo**
Valor Previsto	**Negativo**	True Negativo	False Negativo
	Positivo	False Positive	True Positive

Fonte: do Autor.

As métricas utilizadas para analisar a classificação são a Acurácia (do inglês *Accuracy*), Precisão (do inglês *Precision*) e a Revocação (do inglês *Recall*). Elas podem ser descobertas usando respectivamente as seguintes fórmulas 4.1, 4.2, 4.3

A acurácia é o número de pontos de dados previstos corretamente de todos os pontos de dados. Seguindo formalmente essa fórmula 4.1

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

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. Seguindo formalmente essa fórmula 4.2

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.

$$P = \frac{TP}{TP + FP} \tag{4.2}$$

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. Seguindo formalmente essa fórmula 4.3

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.

$$R = \frac{TP}{TP + FN} \tag{4.3}$$

A 4.2 mostra os valores médios e desvios-padrão dos *folds* encontrados.

Tabela 4.2 – Tabela de 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

Fonte: do Autor.

Tabela 4.3 – Tabela de 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

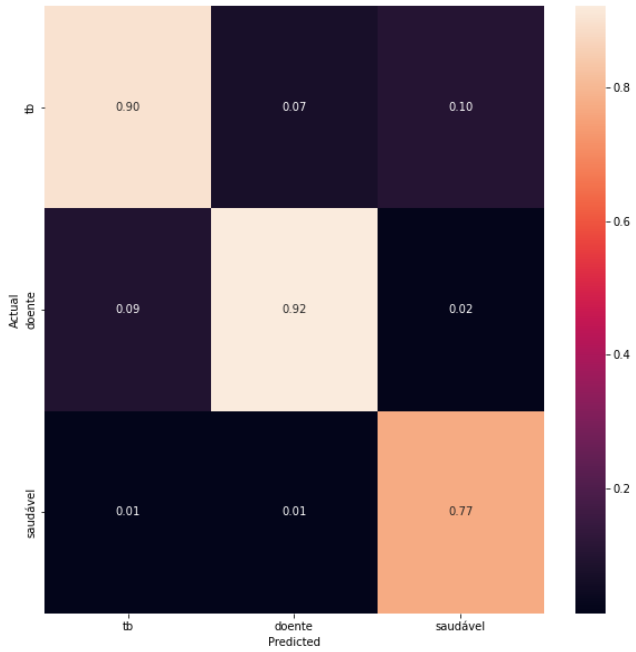
Fonte: do Autor.

Com base no que foi apresentado 4.3, nota-se que o método tem uma acurácia de 90%. O *Precision* e *Recall* mostra 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 melhores que os resultados de teste, indicando boa generalização.

As imagens das matrizes de confusão foram geradas usando a biblioteca *Seaborn* (SEABORN. . . , 2022) para melhor visualização. Observando as figuras 4.1 e 4.2, 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. E é possível observar que por causa da restrição na quantidade de imagens no banco de dados para a classe TB, as matrizes apontam que a classe TB ainda é a classe que mais sofre com a classificação errada, onde imagens de TB são confundidas com imagens de pulmões sadios e doentes.

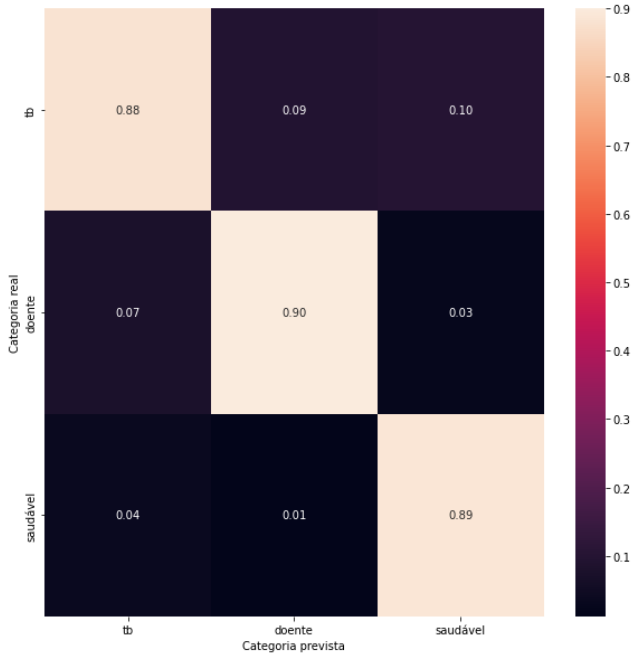
Pelo fato de que as operações feitas pela classificação só calculam a distância da amostra para a curva e calcula e monta as curvas em si, isso simplifica o programa deixando ele com uma complexidade menor e com um desempenho melhor do que outros métodos..

Figura 4.1 – Matriz de Confusão dos valores de precisão da média das CPs normalizada usando Seaborn



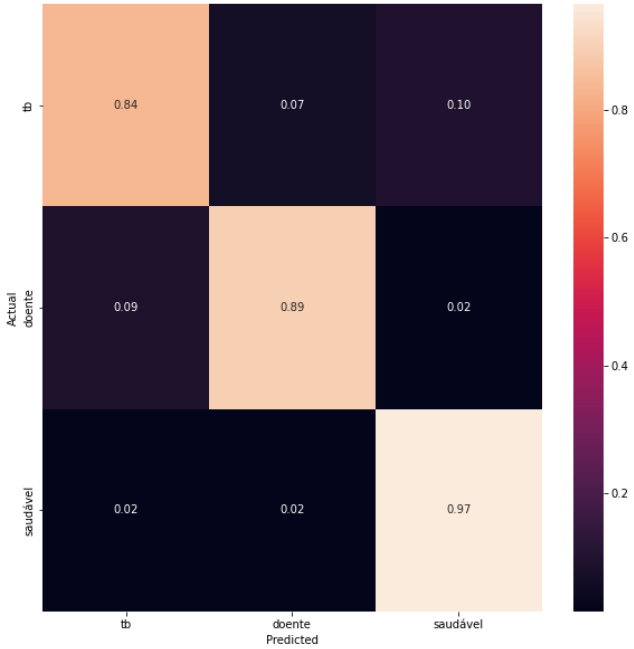
Fonte: do Autor.

Figura 4.2 – Matriz de Confusão dos valores de precisão da melhor CP normalizada usando Seaborn



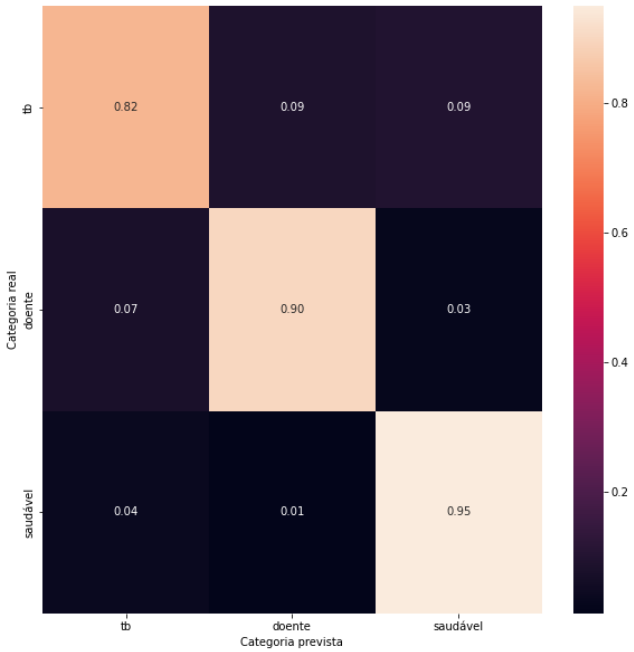
Fonte: do Autor.

Figura 4.3 – Matriz de Confusão dos valores de revocação da média das CPs normalizada usando Seaborn



Fonte: do Autor.

Figura 4.4 – Matriz de Confusão dos valores de revocação da melhor CP normalizada usando Seaborn



Fonte: do Autor.

5 CONCLUSÃO

Foi possível validar, por meio dos resultados do presente trabalho, 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, para que estes problemas se tornem menos limitantes para pesquisadores.

Outro aspecto pertinente visto ao longo desta pesquisa é que a TB, mesmo tão letal e 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.

Quando for comparar esse método com os outros métodos, como por exemplo, a CNN, é necessário não apenas observar no quesito acurácia, mas também no âmbito da exigência computacional do hardware, pois, como explicado previamente, a tecnologia predominantemente usada na área é a CNN, mas ainda não existem estudos conclusivos da viabilidade desse modelo em sistemas *IoT* (Internet of Things), embarcados e em aplicativos de *smartphones*, dessa maneira esse modelo utilizando Curvas Principais com uma exigência de hardware muito inferior que a CNN tem uma vantagem nessas áreas.

REFERÊNCIAS

ALZUBAIDI, L. et al. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. **Journal of Big Data**, v. 8, 03 2021.

AYYACHAMY, S. et al. Medical image retrieval using Resnet-18. In: CHEN, P.-H.; BAK, P. R. (Ed.). **Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications**. SPIE, 2019. v. 10954, p. 1095410. Acesso em: 20 de Setembro de 2022. Disponível em: <https://doi.org/10.1117/12.2515588>.

"BORGES, L. M. et al. 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.

CANTRELL, S. **Top 3 Most Popular Neural Networks**. 2018. Acesso em: 20 de Setembro de 2022. Disponível em: <https://www.excella.com/insights/top-3-most-popular-neural-networks>.

CORNISH, L. **Interactive: Who's funding the COVID-19 response and what are priorities**. 2021. Acesso em: 20 de Setembro de 2022. Disponível em: <https://www.devex.com/news/interactive-who-s-funding-the-covid-19-response-and-what-are-the-priorities-96833>.

CORTEZ, A. O. et al. Tuberculosis in brazil: one country, multiple realities. **Jornal brasileiro de pneumologia : publicacao oficial da Sociedade Brasileira de Pneumologia e Tisilogia**, v. 47, n. 2, p. e20200119, 2021. ISSN 1806-3713. Disponível em: <https://europepmc.org/articles/PMC8332839>.

FERNANDEZ, H. L. **Classificação de Navios Baseada em Curvas Principais**. 12–47 p. Dissertação (Mestrado) — Engenharia de Sistemas e Automação - Universidade Federal do Rio de Janeiro, 2005.

FURTADO, A. et al. A light deep learning algorithm for ct diagnosis of covid-19 pneumonia. **Diagnostics (Basel, Switzerland)**, v. 12, 06 2022.

HASTIE, T.; STUETZLE, W. Principal curves. **Journal of the American Statistical Association**, Taylor & Francis, v. 84, n. 406, p. 502–516, 1989. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478797>.

HUSSAIN, M.; BIRD, J.; FARIA, D. A study on cnn transfer learning for image classification. In: . [S.l.: s.n.], 2018.

IBM. **IBM Watson is AI for smarter business**. 2022. Acesso em: 20 de Setembro de 2022. Disponível em: <https://www.ibm.com/watson>.

LIU, Y. et al. Rethinking computer-aided tuberculosis diagnosis. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2020. p. 2646–2655.

MACIEL, E. L. et al. Tuberculosis: a deadly and neglected disease in the covid-19 era. **Jornal brasileiro de pneumologia : publicacao oficial da Sociedade Brasileira de Pneumologia e Tisilogia**, v. 48, n. 3, p. e20220056, June 2022. ISSN 1806-3713. Disponível em: <https://doi.org/10.36416/1806-3756/e20220056>.

MORAES, E. C. C. **Método não supervisionado baseado em curvas principais para reconhecimento de padrões**. Dissertação (Mestrado) — Universidade Federal de Lavras, 2015.

PAI TEREZA KASAEVA, S. S. M. Covid-19's devastating effect on tuberculosis care — a path to recovery. **The New England Journal of Medicine**, 01 2022.

PENG, T. et al. Detection of lung contour with closed principal curve and machine learning. **Journal of Digital Imaging**, v. 31, 02 2018.

PETERSON TERENCE TOLAND, G. H. E. R. **Robots vs. COVID-19: how the pandemic is accelerating automation**. 2020. <https://www.kearney.com/web/global-business-policy-council/article/-/insights/robots-vs-covid-19-how-the-pandemic-is-accelerating-automation>.

"ROCHA, G. G. M. d.; FILHO, J. B. d. O. e. S. Classificação de contatos baseada em curvas principais utilizando o processador nios ii. **Congresso Brasileiro de Automática**, UFMG, SBelo Horizonte, MG, 09 2014.

SEABORN: statistical data visualization. 2022. Acesso em: 20 de Setembro de 2022. Disponível em: <https://seaborn.pydata.org/>.

STEGER, A. **Healthcare Automation Matters More Than Ever During a Pandemic**. 2020. <https://healthtechmagazine.net/article/2020/07/healthcare-automation-matters-more-ever-during-pandemic-perfcon>.

VERBEEK, J.; VLASSIS, N.; KRöSE, B. A k-segments algorithm for finding principal curves. **Pattern Recognition Letters**, v. 23, n. 8, p. 1009–1017, 2002. ISSN 0167-8655. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167865502000326>.

WANG, K. et al. Pay attention to features, transfer learn faster cnns. In: **International Conference on Learning Representations**. [s.n.], 2020. Acesso em: 20 de Setembro de 2022. Disponível em: <https://openreview.net/forum?id=ryxyCeHtPB>.

WATSON, I. **O Watson Health é a saúde mais inteligente**. 2022. Acesso em: 20 de Setembro de 2022. Disponível em: <https://www.ibm.com/br-pt/watson-health>.

ZAMAN, K. Tuberculosis: a global health problem. **Journal of health, population, and nutrition**, v. 28, n. 2, p. 111—113, April 2010. ISSN 1606-0997. Disponível em: <https://europepmc.org/articles/PMC2980871>.