**On the limits of cross-domain generalization in automated X-ray prediction**

<https://arxiv.org/pdf/2002.02497.pdf>

Este trabalho busca evidenciar os problemas relacionados à deliberada utilização de bancos de dados para treino de CNNs para diagnóstico. É apresentado o contraste entre performance e agreement (Cohen’s Kappa\*).

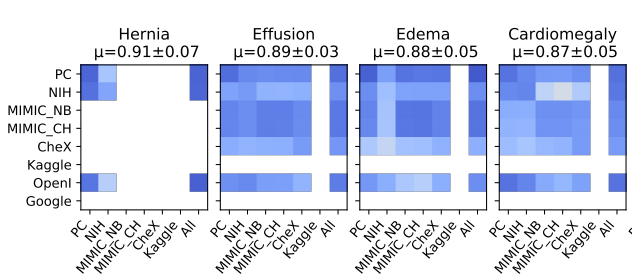
\*Métrica que quantifica o real acerto das redes, acurácia, em relação à concordância randómica entre dois ou mais avaliadores, neste caso, datasets. (<https://en.wikipedia.org/wiki/Cohen%27s_kappa>)

O artigo parte do princípio de que há dissonância entre as categorizações das classes dos dataset, devido há discordância entre radiologistas, clínicos e máquinas, erros em classificadores automáticos e diferenças entre contextos dos classificadores, seja pela língua, cultura médica ou política.

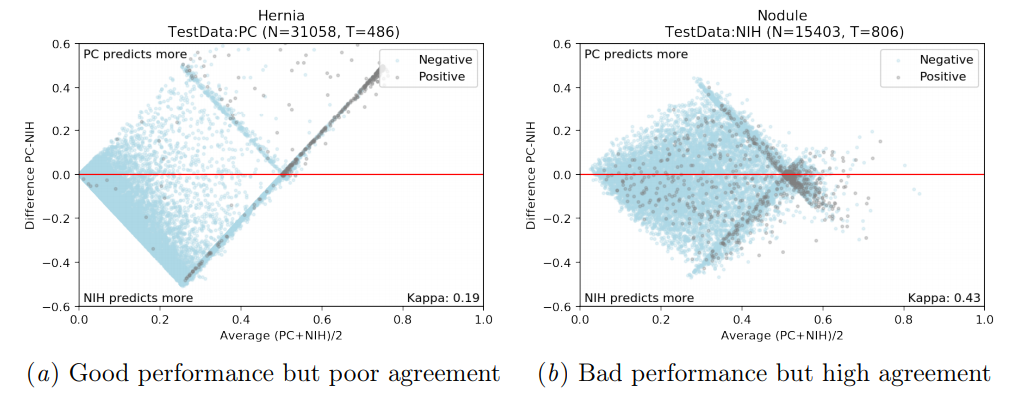
Sete datasets são utilizados. Cada experimentação conta com três redes, o resultado final será a média da saída destas, assim, reduz-se os erros por ruídos.

Os autores contornam o problema de super-representação de determinadas classes ajustando a função de perda, baseando-se no balanço dos pesos pela proporção de cada classe. As fórmulas estão presentes no artigo. Em seguida, o melhor threshold é obtido e utilizado para as experimentações posteriores, que não é o mesmo para todos os datasets.

Para a avaliação de performance, a rede é treinada em todas os bancos, um por vez, e depois em todos, depois, testada nos bancos excludentes, uma classe por vez. Isso gera um heatmap de ROC-AUC por cada classe. Assim, é possível analisar qual banco de dados é melhor para treino com o objetivo de classificar determinado outro banco em uma determinada classe. Um exemplo segue abaixo.



A escala vai do vermelho ao azul escuro, em ordem crescente, de 0 a 1. Os blocos brancos significam que o banco não continha a classe.

Em seguida, o agreement entre a rede treinada em um dataset, comparada as redes treinadas em outros datasets, referentes à mesma classe é analisado. Nota-se que em alguns casos há alta AUC, mas há, também, disagreement (baixo Kappa). Isso significa que o coeficiente de fator randómico está muito próximo à acurácia, ou seja, há diferenças entre a classe em questão entre os diferentes bancos (os possíveis motivos foram citados no começo do texto). O gráfico abaixo, Bland Altman plot, exemplifica os casos.

Particularmente, ainda não consegui interpretar o plot totalmente. Apenas sei que no caso ideal, os dados devem estar dispostos o mais próximo e uniformemente possível da linha vermelha.

Outra análise é realizada, a saída da rede é modificada para que cada dataset tenha a saída de todas suas classes, totalizando 90 saídas. Dois modelos são treinados. Após o treinamento, PCA é realizado com pesos de cada classe, gerando um vetor 2d, e a rede é treinada novamente, desta vez, com a inserção dos vetores, forçando-os a se alinharem. No caso ideal, os vetores teriam direção, sentido e magnitude muito próximos no fim do processo, mas não é o caso, ainda há classes em que os vetores destoam bastante, evidenciando as diferenças entre os datasets.

Com todos esses resultados, os autores concluem que os modelos visando diagnóstico não devem apenas se limitar em melhorar a performance, covariância, mas também nas evidencias apresentadas, agreement entre os bancos.

Certamente eu não consegui sintetizar o conteúdo, há muita coisa fora dos meus conhecimentos que preciso amadurecer, convido-te a ler o artigo original usando como ponto de partida meu texto.