2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-1179-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/EMBC46164.2021.9630898

Detecting COVID-19 Related Pneumonia On CT Scans Using Hyperdimensional Computing

Neftali Watkinson[1], Tony Givargis[1], Victor Joe[2], Alexandru Nicolau[1] and Alexander Veidenbaum[1]

Abstract— **Pneumonia is a common complication associated with COVID-19 infections. Unlike common versions of pneumonia that spread quickly through large lung regions, COVID-19 related pneumonia starts in small localized pockets before spreading over the course of several days. This makes the infection more resilient and with a high probability of developing acute respiratory distress syndrome. Because of the peculiar spread pattern, the use of pulmonary computerized tomography (CT) scans was key in identifying COVID-19 infections. Identifying uncommon pulmonary diseases could be a strong line of defense in early detection of new respiratory infection-causing viruses. In this paper we describe a classification algorithm based on hyperdimensional computing for the detection of COVID-19 pneumonia in CT scans. We test our algorithm using three different datasets. The highest reported accuracy is 95.2% with an F1 score of 0.90, and all three models had a precision of 1 (0 false positives).**

I. Introduction

On January 2020, the World Health Organization declared a global emergency due to a novel coronavirus that had started in the regions of Wuhan, China and rapidly spread around the world. Symptoms include fever, headache, loss of smell, difficulty breathing, among others [1]. During the early stages of infection, medical experts were specifically interested in the effects that SARS-CoV-2 (COVID-19) had in the patient's lungs. In severe cases, lungs get inflamed, filled with fluid and debris, causing what is known as pneumonia [2]. Because of this, hospitals relied on computerized tomography (CT) scans [3] and using lower respiratory tract samples [4] to identify COVID-19 related pneumonia. Accuracy among expert radiologist on differentiating COVID-related pneumonia from typical pneumonia can vary fell between 97% and 67% [5].

At the time of writing, various blood and saliva tests have been developed, but their accuracy has been the focus of debate. While most tests seem to emphasize sensitivity over specificity [6], CT scans remain the most accurate way to confirm symptomatic infections, especially when dealing with in-hospital settings and with patients with preexisting lung related conditions [7].

Recent research has focused on applying artificial intelligence to the challenge of detecting COVID-19 in CT scans [8]. The work of Soares, et al. [9] aims to build an explainable deep learning network using CT scans collected from patients across several hospitals in Sao Paulo, Brazil. They released a dataset containing images from healthy patients, patients with COVID-19 related pneumonia, and patients with other pulmonary issues. Similarly, He, et al. [10] released a dataset containing ct-scans of healthy patients and patients diagnosed with COVID-19 that were data mined from other studies. Finally, Rahimzadeh, et. al. [11] released a dataset of CT scan sequences sourced from a single hospital, and introduced a deep learning based model to identify COVID-19 infections. In total, we have access to three public datasets.

In this paper we describe a hyperdimensional (HD) computing [12] approach for identifying CT scan images as suspicious of COVID-19 using image classification. We train and test the models using each of the three datasets. Our approach achieves up to 95% classification accuracy and report 0 false positives. We compared our models with those originally published alongside the datasets and present a thorough discussion on the advantages and disadvantages of our approach.

The rest of the paper is organized as follows:

- **Methodology** explains our HD computing implementation and justifies decisions made during the design of the classification model.
- **Results** describes the resulting accuracy with each of the CT scans datasets and compares them with the models described in the original papers that published the datasets.
- In **Discussion** we justify our approach and describe its differences over other approaches.
- **Conclusion** summarizes our findings.

II. Methodology

Our algorithm relies on HD computing to encode one gray scale image sample per CT scan into a *hypervector* (a very long vector of numbers). For this work we use vectors with 10 thousand elements. After the data is encoded, we can compute the distance between hypervectors. The predicted class for unknown encoded images will be that of the closest known hypervector. In this section we describe the specific characteristics of our implementation.

A. HD Computing for Image Classification

This is not the first work that uses HD computing for image classification [13]–[15]. There different implementations, but we follow an algorithm very similar to the one described by Yang, et al. [16]. We use bipolar hypervectors

[1] Neftali Watkinson, Tony Givargis, Alexandru Nicolau and Alexander Veidenbaum are with the Donald Bren School of Information and Computer Sciences, University of California, Irvine Donald Bren Hall, 6210, Irvine, CA 92697 watkinso@uci.edu, givargis@uci.edu, nicolau@ics.uci.edu, alexv@ics.uci.edu

[2] Victor Joe is with the Regional Burn Center, UCI Medical Center, 101 The City Dr S, Orange, CA 92868 vcjoe@hs.uci.edu

(every element has a value of 1 or -1) and encode both the intensity and the position of every pixel. After obtaining both hypervectors, we bind them through multiplication. Then we combine all the pixel hypervectors through majority voting [12] to generate a single hypervector for the image.

B. Encoding

For creating the hypervectors, we use orthogonal or uncorrelated encoding [13]–[15] to represent the position of each pixel. We use linear or correlated encoding [17] to represent the pixel intensity. This means that each pixel has two hypervectors, which are later combined through majority voting.

1) Preprocessing: Through the use of image manipulation tools, we resized the images and normalized their contrast ratio to remove noise introduced from different sources. Preprocessing and image filtering is not unusual in machine learning-based image processing tasks [18]–[20], and commonly used for face [21] and object [22] detection.

2) Training: Once all the images are preprocessed, and the dimensionality (10k) and type (bipolar) of the hypervectors has been set, the training phase proceeds as follows:

1) Identify the magnitude of features. Since all the images are the same size after processing, then we have 300 by 200 pixels or 60,000 features for datasets 1 [9] and 2 [10], and 512 by 512 (262,144) pixels for dataset 3 [11].
2) For each pixel, encode the position and the intensity hypervector and bind them.
3) Combine all the pixel hypervectors of a single image using majority voting.
4) Add all the feature vectors using majority voting.
5) Store the image hypervector keeping record of what class (healthy or COVID-19) it belongs to.

3) Testing: We separate a subset of the images that are not part of the training phase. This subset is called the testing subset and is used to evaluate the model. For this work we use a 70-30 split, which means that 70% of the images will be used for training and the remaining 30% will be used for testing. They all go through the same encoding process with the only difference that the testing images' class is set to that of the closest training hypervector. We then compare this class with the actual class of the original image and derive the accuracy of the model in correctly predicting the testing images.

III. Results

Imaging data from CT scans are generally stored using the DICOM formatting [23] that contains information such as patient data. For the three datasets, the images have been scrubbed of identifying data and extracted as Portable Network Graphics (PNG) images. Axial CT scans are done from the perspective of the axial or transverse plane, along or perpendicular to the median plane. In other words, with the patient lying on their back, slices are collected starting from the upper lobe of the lungs (closest to the patient's head) towards the lower lobe (closest to the patient's waist). Figure 1 shows an image sample for a single patient. Ranges and slice sizes vary for each patient. We chose to focus on the mid section with the right major fissure in focus and the trachea splitting into the main bronchi (b) since this was fairly consistent among all patients for all datasets.

We built three models, one for each dataset. From dataset 1, we have 80 images from patients with COVID-19 and 46 images from healthy patients. From dataset 2 [10], we have 350 with COVID-19 and 398 without. From dataset 3, we used one image per patient from all 96 patients with COVID-19 and randomly sampled 109 healthy patients in order to keep the dataset balanced. Table I shows the population distribution for each dataset and Figures 2 and 3 show an image sample of a healthy and a COVID-19 CT scan image respectively.

	COVID-19	Healthy	Total
Dataset 1	80	46	126
Dataset 2	350	398	748
Dataset 3	96	109	205

TABLE I

THIS TABLE SHOWS THE POPULATION DISTRIBUTION FOR EACH OF THE DATASETS

Model	Accuracy	Precision	Recall	F1 score
Model 1	92.8%	**1**	0.83	0.91
xDNN	**97.38%**	0.99	**0.95**	**0.97**
Model 2	**93.7%**	1	0.77	**0.87**
Self-Trans	86%	-	-	0.85
Model 3	95.2%	**1**	0.82	**0.91**
FPN	**98.5%**	0.73	**0.94**	0.82

TABLE II

ACCURACY, PRECISION, RECALL AND F1 SCORE FOR OUR MODELS COMPARED TO THE MODEL USED FOR EACH OF THE PUBLISHED DATASETS IN THEIR ORIGINAL PAPERS. MODEL 1 WAS USED FOR DATASET 1, MODEL 2 FOR DATASET 2 AND MODEL 3 FOR DATASET 3. FPN STANDS FOR FEATURE PYRAMID NETWORK.

A. Comparing to expert radiologists

For the models corresponding to each dataset we derived the classification accuracy, the precision, recall and F1 score and compared them to the models used in the original papers for each dataset. These results are in Table III. It is important to note that the model originally presented in the work of Soares, et al. [9] has a data leakage issue where images where randomly split without separating them per patient. This means that images from the same patient can appear in the training and the validation set. All of our models had a precision of 1 which means that there were no false positives, the reasoning for this is that the haziness in images from infected lungs introduces pixel values at specific image locations that are easily identified by the model. A healthy and clear lung will have pixel values closer to absolute white or absolute black (255 or 0 correspondingly) whereas the haze in infected lungs will have pixel values closer to the middle of the pixel intensity range.

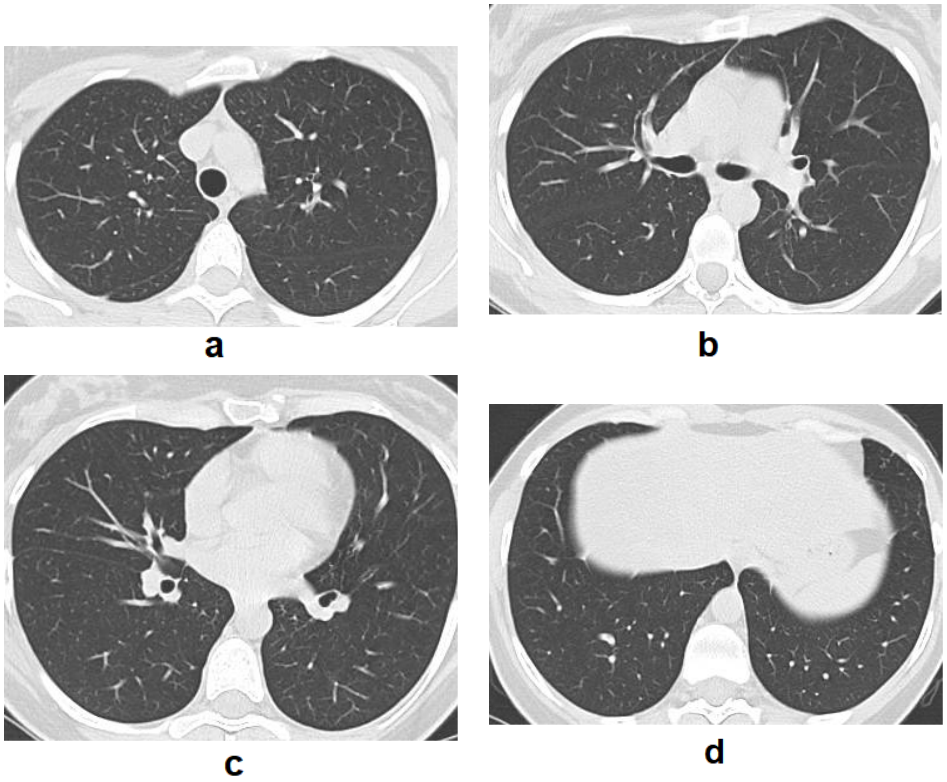

Fig. 1. CT scan sample from one patient showing the upper lobe and the trachea in the middle of the image(a), the mid section of the lung showing the anterior segment in focus and the trachea splitting into the main bronchi (b), the lower lobe section with inferior lobar bronchi (c), and the basal segments of the lower lobe with the diaphragm starting to appear (d)

Fig. 2. Image from a CT scan of a patient with healthy lungs showing minimal haziness and clear definition of the arteries

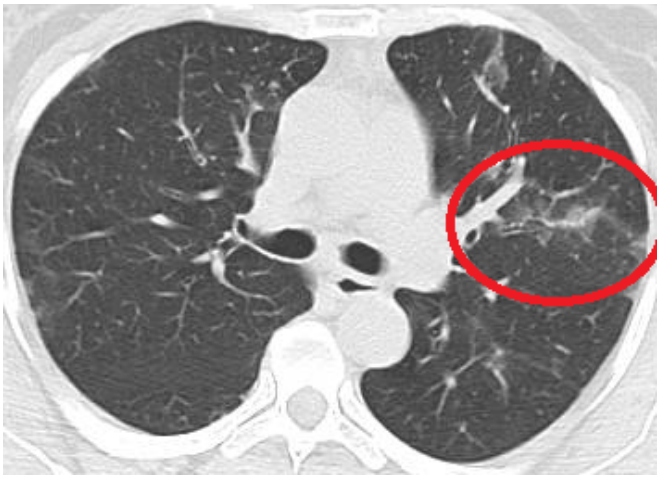

Fig. 3. Image from a CT scan of a patient with COVID-19. The haziness (as the one found inside the oval) is indicative of pneumonia and the pattern is consistent with that of patients with a COVID-19 infection.

IV. Comparing to expert radiologists

The classification performance of the binary classification model is comparable to the median accuracy values presented in [5]. However, there are key differences between the two studies. In [5], CT scans from COVID-19 patients with no abnormalities were discarded. Additionally, radiologist had access to the full scan. For the model presented in this chapter, only one slice is being analyzed and all COVID-19 are included, without discarding non-anomalous images. [11] observed in a separate experiment that radiologists where only 70% accurate in detecting COVID-19 infections in CT scans. Future research needs to be done to discover an encoding that is not dependent upon pixel position and that could be implemented to three-dimensional images.

V. Discussion

HD computing has proven to be an efficient machine learning approach for many domains, specially those that are

data constrained [13], [17], [24], [25]. In this work we show that classification of CT scan images is another use for it, effectively capturing the pixel patterns observed in CT scans of lungs with a COVID-19 infection. Our models consistently achieve over 92% classification accuracy and surpasses deep learning based models in at least one of the metrics. An advantage that HD computing has that we don't discuss in this paper is that it is computationally efficient [12].

On the other hand, our HD computing models perform well with little to medium amounts of data. Dataset 1 only contains 126 images and dataset 2 contains 748. This means that this approach has both the potential of being used to detect new unseen anomalies where data is scarce, or scale and improve as more data is made available. However, there is still much work to do in order to properly assess this, in addition to incorporating a full CT scan sequence as a hypervector, with the option of adding the patient's data as well.

For future works we will focus on testing the impact of data scarcity in identifying anomalous CT scans. This is part of a greater effort to design tools that will help identify new respiratory diseases when little data is available in support of an early detection system.

VI. Conclusion

In this paper we describe a hyperdimensional computing image classification approach to classifying images from pulmonary CT scans across two classes: Healthy and with COVID-19 related pneumonia. We test our approach through three models, each one applied to a different dataset. All the models achieve over 92% classification accuracy and beat the state of the art models that were used originally on these datasets by at least one metric (accuracy, precision, recall and F1 score). Additionally, none of the models generate any false negatives and greatly reduce the input dimensionality. We argue that this approach has the potential of detecting new pulmonary diseases but also scales well when more data is made available.

The medical data used in this study was fully de-identified and ethical committees relevant to each of the referenced sources provided approval to the original works. For this paper, we did not perform experimental procedures on humans and followed the data reuse guidelines provided by the authors of the original data sources.

References

[1] A. Carfì, R. Bernabei, F. Landi *et al.*, "Persistent symptoms in patients after acute covid-19," *Jama*, vol. 324, no. 6, pp. 603–605, 2020.

[2] F. Pan, T. Ye, P. Sun, S. Gui, B. Liang, L. Li, D. Zheng, J. Wang, R. L. Hesketh, L. Yang *et al.*, "Time course of lung changes on chest ct during recovery from 2019 novel coronavirus (covid-19) pneumonia," *Radiology*, 2020.

[3] P. Huang, T. Liu, L. Huang, H. Liu, M. Lei, W. Xu, X. Hu, J. Chen, and B. Liu, "Use of chest ct in combination with negative rt-pcr assay for the 2019 novel coronavirus but high clinical suspicion," *Radiology*, vol. 295, no. 1, pp. 22–23, 2020.

[4] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu *et al.*, "A novel coronavirus from patients with pneumonia in china, 2019," *New England Journal of Medicine*, 2020.

[5] H. X. Bai, B. Hsieh, Z. Xiong, K. Halsey, J. W. Choi, T. M. L. Tran, I. Pan, L.-B. Shi, D.-C. Wang, J. Mei *et al.*, "Performance of radiologists in differentiating covid-19 from viral pneumonia on chest ct," *Radiology*, p. 200823, 2020.

[6] R. Castro, P. M. Luz, M. D. Wakimoto, V. G. Veloso, B. Grinsztejn, and H. Perazzo, "Covid-19: a meta-analysis of diagnostic test accuracy of commercial assays registered in brazil," *The Brazilian Journal of Infectious Diseases*, 2020.

[7] B. Xu, Y. Xing, J. Peng, Z. Zheng, W. Tang, Y. Sun, C. Xu, and F. Peng, "Chest ct for detecting covid-19: a systematic review and meta-analysis of diagnostic accuracy," *European Radiology*, p. 1, 2020.

[8] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song *et al.*, "Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct," *Radiology*, 2020.

[9] E. Soares, P. Angelov, S. Biaso, M. H. Froes, and D. K. Abe, "Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification," *medRxiv*, 2020.

[10] X. He, X. Yang, S. Zhang, J. Zhao, Y. Zhang, E. Xing, and P. Xie, "Sample-efficient deep learning for covid-19 diagnosis based on ct scans," *medrxiv*, 2020.

[11] M. Rahimzadeh, A. Attar, and S. M. Sakhaei, "A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset," *Biomedical Signal Processing and Control*, p. 102588, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421001853

[12] P. Kanerva, "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors," *Cognitive computation*, vol. 1, no. 2, pp. 139–159, 2009.

[13] D. Kleyko, A. Rahimi, D. A. Rachkovskij, E. Osipov, and J. M. Rabaey, "Classification and recall with binary hyperdimensional computing: Tradeoffs in choice of density and mapping characteristics," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 12, pp. 5880–5898, 2018.

[14] G. Recchia, M. Sahlgren, P. Kanerva, and M. N. Jones, "Encoding sequential information in semantic space models: Comparing holographic reduced representation and random permutation," *Computational intelligence and neuroscience*, vol. 2015, 2015.

[15] M. Sahlgren, A. Holst, and P. Kanerva, "Permutations as a means to encode order in word space," in *The 30th Annual Meeting of the Cognitive Science Society (CogSci'08), 23-26 July 2008, Washington DC, USA*, 2008.

[16] F. Yang and S. Ren, "On the vulnerability of hyperdimensional computing-based classifiers to adversarial attacks," in *International Conference on Network and System Security*. Springer, 2020, pp. 371–387.

[17] A. Rahimi, P. Kanerva, L. Benini, and J. M. Rabaey, "Efficient biosignal processing using hyperdimensional computing: Network templates for combined learning and classification of exg signals," *Proceedings of the IEEE*, vol. 107, no. 1, pp. 123–143, 2018.

[18] T. Bangira, S. M. Alfieri, M. Menenti, and A. Van Niekerk, "Comparing thresholding with machine learning classifiers for mapping complex water," *Remote Sensing*, vol. 11, no. 11, p. 1351, 2019.

[19] J. Ker, S. P. Singh, Y. Bai, J. Rao, T. Lim, and L. Wang, "Image thresholding improves 3-dimensional convolutional neural network diagnosis of different acute brain hemorrhages on computed tomography scans," *Sensors*, vol. 19, no. 9, p. 2167, 2019.

[20] A. Chowdhury, E. Kautz, B. Yener, and D. Lewis, "Image driven machine learning methods for microstructure recognition," *Computational Materials Science*, vol. 123, pp. 176–187, 2016.

[21] D. Bradley and G. Roth, "Adaptive thresholding using the integral image," *Journal of graphics tools*, vol. 12, no. 2, pp. 13–21, 2007.

[22] M. P. De Albuquerque, I. A. Esquef, and A. G. Mello, "Image thresholding using tsallis entropy," *Pattern Recognition Letters*, vol. 25, no. 9, pp. 1059–1065, 2004.

[23] P. Mildenberger, M. Eichelberg, and E. Martin, "Introduction to the dicom standard," *European radiology*, vol. 12, no. 4, pp. 920–927, 2002.

[24] M. Imani, D. Kong, A. Rahimi, and T. Rosing, "Voicehd: Hyperdimensional computing for efficient speech recognition," in *2017 IEEE International Conference on Rebooting Computing (ICRC)*. IEEE, 2017, pp. 1–8.

[25] M. Imani, S. Salamat, S. Gupta, J. Huang, and T. Rosing, "Fach: Fpga-based acceleration of hyperdimensional computing by reducing computational complexity," in *Proceedings of the 24th Asia and South Pacific Design Automation Conference*, 2019, pp. 493–498.