Five Strategies for Bias Estimation in Artificial Intelligence-based Hybrid Deep Learning for Acute Respiratory Distress Syndrome COVID-19 Lung Infected Patients using AP(ai)Bias 2.0: A Systematic Review

Jasjit S. Suri, *Fellow IEEE*, Sushant Agarwal, *Member, IEEE*, Biswajit Jena, Sanjay Saxena, Member, IEEE, Ayman El-Baz, Vikas Agarwal, Mannudeep K. Kalra, Luca Saba, Klaudija Viskovic, Mostafa Fatemi, *Life Fellow, IEEE*, Subbaram Naidu, *Life Fellow, IEEE*

Abstract—**Coronavirus 2019 (COVID-19) has led to a global pandemic infecting 224 million people and has caused 4.6 million deaths. Nearly 80 Artificial Intelligence (AI) articles have been published on COVID-19 diagnosis. The first systematic review on the Deep Learning (DL)-based paradigm for COVID-19 diagnosis was recently published by Suri *et al*. [IEEE J Biomed Health Inform. 2021]. The above study used AtheroPoint's "AP(ai)Bias 1.0" using 10 AI attributes in the DL framework.**

The proposed study uses "AP(ai)Bias 2.0" as part of the three quantitative paradigms for *Risk-of-Bias* quantification by using the best 40 dedicated Hybrid DL (HDL) studies and utilizing 39 AI attributes. In the *first method*, the *radial-bias* map (RBM) was computed for each AI study, followed by the computation of bias value. In the *second method*, the *regional-bias* area (RBA) was computed by the area difference between the best and the worst AI performing attributes. In the *third method*, *ranking-bias* score (RBS) was computed, where AI-based cumulative scores were computed for all the 40 studies. These studies were ranked, and the cutoff was determined, categorizing the HDL studies into three bins: low, moderate, and high. Using the Venn diagram, these three quantitative methods were benchmarked against the two qualitative non-randomized-based AI trial methods (ROBINS-I and PROBAST).

Using the analytically derived moderate-high and low-moderate cutoff of 2.9 and 3.6, respectively, we observed 40%, 27.5%, 17.5%, 10%, and 20% of studies were *low-bias*ed for RBM, RBA, RBS, ROBINS-I, and PROBAST, respectively. We present an eight-point recommendation for AP(ai)Bias 2.0 minimization.

Index Terms—**COVID-19 diagnosis, HDL, risk-of-bias, radial-regional-ranking, PROBAST-ROBINS-I, AP(ai)Bias 2.0.**

I. INTRODUCTION

Since the outbreak of novel coronavirus (SARS-CoV-2) in December 2019, the world health organization (WHO) has declared it as a global pandemic [1], called COVID-19. As per the WHO's statistical records, this deadly disease has infected **224 million** people had caused **4.6 million** deaths across the globe [2]. The world is still reeling under this deadly virus, with several waves of disease noticed in the form of mutants such as alpha, beta, and, recently, delta [3]. The ongoing COVID-19 situation has put the healthcare sector in more vulnerable situations, causing intense pressure on the pharmaceutical and financial sectors [4].

Real-time reverse transcription-polymerase chain reaction (RT-PCR) is the standard clinician's strategy to discover the presence or absence of this type of virus [5]. RT-PCR has a relatively low positive rate and sensitivity for the early detection of this disease [6-9]. The SAR-CoV-2 affects various organs of the body such as the lungs, coronary artery, carotid artery [10-12], and brain [13, 14], causing pulmonary embolism [15] through the molecular pathways [16], accelerating diabetes [17], and leading to cerebral thrombosis [18]. X-ray and Computed Tomography (CT)-based lung imaging is used as an alternative for understanding the severity of the disease, particularly in the quantification of ground-glass opacities (GGO) in CT scans [19-22]. Manual methods by radiologists are used for judging the COVID-19 severity, but it is tedious, slow, vulnerable to errors, and yields low specificity and sensitivity [6-9]. Thus, there is a need for a reliable, automated, scientific and clinically validated, real-time solution for the early COVID-19 disease diagnosis and prognosis, thereby saving human lives.

Artificial Intelligence (AI) has penetrated all walks of life and, more recently, in the field of imaging sciences [23], particularly in big data frameworks [24]. These AI-based studies that have emerged during COVID-19 demonstrate exceptionally high performance without strong clinical outcomes and therefore are considered biased [25-27]. The team of immunologists, pulmonologists, radiologists, and cardiologists are interested in using the AI technology to diagnose COVID-19 severity, but it has produced a dent that has gone unnoticed several times due to AI bias [28]. The proposed study is focused on methods to quantify the biased nature of AI-based solutions for COVID-19 diagnosis. Recently, the first systematic review on the deep learning (DL) paradigm for COVID-19 diagnosis was published by Suri *et al.* [28]. In this study, AtheroPoint's "**AP(ai)Bias 1.0**" was designed using **10** AI attributes in the **DL** framework based on the ranking paradigm. Even though the study was innovative, it is not robust enough to handle a large number of AI attributes. Since HDL is combination of two solo DL (SDL) models, either cascaded [29] or parallel [30], therefore, these configurations can affect the performance of the segmentation or classification. Thus, it is important to study the bias in HDL.

JSS is with Stroke Diagnosis and Monitoring Division, AtheroPoint™, Roseville, CA, USA. SA is with Advanced Knowledge Engineering Centre, Global Biomedical Technologies, Inc., Roseville, CA, USA and Dept. of CSE, PSIT, Kanpur, UP, INDIA. BJ and SS is with Dept. of CSE, International Institute of Information Technology, Bhubaneswar, INDIA. AE is with Dept. of Bioengineering at the University of Louisville, KY. VA is with Dept. of Immunology, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, UP, INDIA. MK is with Department of Radiology, Massachusetts General Hospital, 55 Fruit Street, Boston, MA, USA. LS is with Department of Radiology, University of Cagliari, ITLAY. KA is with University Hospital for Infectious Diseases, Zagreb, CROATIA, MF is with Dept. of Physiology & Biomedical Engg., Mayo Clinic College of Medicine and Science, MN, USA. SN is with Electrical Engineering Department, University of Minnesota, Duluth, MN, USA. *Corresponding author: Jasjit S. Suri (jasjit.suri@atheropoint.com)

The proposed study presents *three new innovative paradigms* for bias estimation in the best 40 dedicated HDL studies for COVID-19 diagnosis utilizing 39 (nearly four times) AI attributes. These *three* methods are then benchmarked against earlier two non-randomized-based AI trial methods (ROBINS-I [31] and PROBAST [32]) for further analysis. In the *first* innovative solution, a *radial-bias* map was computed for each AI study, followed by its bias measurement. In the *second* innovative method, the AI-bias of a study was computed by taking the area difference between the *best AI performing attributes* and *least AI performing* attributes. In the *third* method, cumulative scores for all the AI studies were computed and then ranked, dividing the 40 HDL studies into three bins corresponding to *low-bias*, *moderate-bias*, and *high-bias*. These three quantitative and innovative methods, classified as "**AP(ai)Bias 2.0"**, were then benchmarked against the two qualitative non-randomized-based AI trial methods (ROBINS-I and PROBAST). Finally, we use the Venn Diagram (VD) to estimate the total studies common between the five types of AI-bias methods or between any combinatorial AI-bias pairs.

The remaining part of the study has the following layout. Section II presents the search strategy using the PRISMA model. The statistical analysis on various HDL attributes is also presented in the same section. Section III shows the basic principle of the HDL architecture along with the strategy for bias estimation. Section IV presents the three innovative solutions for bias measurement and its interpretation. Section 5 shows the benchmarking strategy using ROBINS-I and PROBAST methods, along with the data analysis, and finally, section VI presents the discussions, followed by the conclusions in section VII.

II. Search Model and statistical distributions

A. The PRISMA Model

A detailed search was performed using Google Scholar, PubMed, IEEE Xplore, ScienceDirect, and arXiv. The keywords used for selecting studies were "hybrid deep learning for COVID", "hybrid deep learning for COVID classification", "hybrid models", "COVID diagnosis using the hybrid model", "hybrid deep transfer learning for COVID classification", and "transfer learning-based deep hybrid model for COVID application". Figure 1 shows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) model flow diagram consisting of the HDL application on COVID diagnosis references used in this study. A total of **183** studies were identified; duplicates were removed leaving 123 records using the "Find Duplicates" feature in EndNote software by Clarivate Analytics. The three exclusion criteria were (**i**) studies not related to AI, (**ii**) non-relevant articles, and (**iii**) articles with insufficient data. After applying the exclusion criteria, 23, 20, and 10 studies (marked as **E1**, **E2**, and **E3** in Figure 1) were identified and removed, leading to the final selection of crucial **70** references for this study.

B. Hypothesis and Risk-of-Bias Acceptability Criteria

We hypothesize that "non-randomized HDL-based attributes can (**a**) detect, (**b**) classify, (**c**) estimate severity of the COVID-19 risks, and (**d**) meet the performance standards in lung infected Acute Respiratory Distress Syndrome (ARDS) patients." The first three out of five acceptability criteria under which the HDL-based studies are considered for bias estimation were: (**i**) *radial-bias* map (RBM) method, (**ii**) *regional-bias* area (RBA) method, (**iii**) *ranking-bias* score (RBS) method, the mean score must be greater than or equal to **80%** for an AI-HDL based study while taking into consideration all the AI-HDL attributes. This was due to the consensus of the experienced team and three different classes of bias for each AI attribute based on its strength (such as low, moderate, and high). Similarly for the remaining two, (**iv**) ROBINS-I and (**v**) PROBAST paradigms, our acceptability criterion must meet the score of **80%** or above for HDL-based studies to be in the low Risk-of-Bias (RoB) zone.

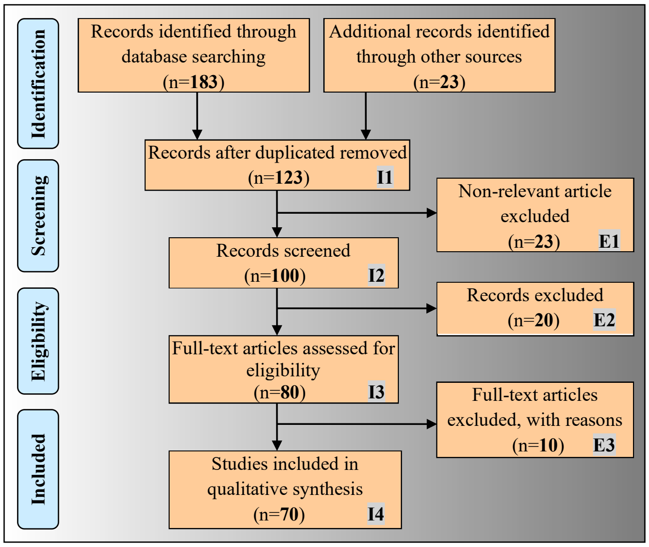

Identification
Screening
Eligibility
Included
Records identified through database searching (n=183)
Additional records identified through other sources (n=23)
Records after duplicated removed (n=123) I1
Non-relevant article excluded (n=23) E1
Records screened (n=100) I2
Records excluded (n=20) E2
Full-text articles assessed for eligibility (n=80) I3
Full-text articles excluded, with reasons (n=10) E3
Studies included in qualitative synthesis (n=70) I4

Figure 1. The PRISMA model.

C. Statistical Distributions by different Criteria

C.1 Statistical Distribution by the Four Types of Objectives

The diagnosis of COVID-19 using the HDL paradigm is classified into four objectives when considering lung CT imaging. (i) Lung classification (**LC**) for COVID-19 detection; (ii) Lung classification followed by lesion localization (**LC+LL**); (iii) Lung segmentation followed by lung classification (**LS+LC**); (iv) Lung segmentation followed by lung classification and lesion localization (**LS+LC+LL**). Using the total of 40 HDL studies, the number of studies under each of these objectives for COVID diagnosis were 26 (**65%**) [6, 20, 33-56], 4 (**10%**) [57-60], 6 (**15%**) [7, 9, 61-64] and 4 (**10%**) [8, 19, 65, 66], respectively (see Figure 2 (a)). Class (i) **(LC)** and Class (ii) **(LC+LL)** used X-ray images or CT images as part of the imaging modality, while Class (iii) (**LS+LC**) and Class (iv) (**LS+LC+LL**) employed CT images as part of lung segmentation as their objective for diagnosis of COVID-19 severity. The studies that provide LL used heatmaps based on Grad-CAM for visualization [19].

C.2 Statistical Distribution by Two Type of Image Modality

Image modality always plays a vital role in the system of COVID-19 diagnosis using HDL paradigms. The two major modalities used in these studies are (a) X-ray and (b) CT. X-ray offers advantages over CT mainly due to low radiation [50],

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2022.3174270, IEEE Transactions on Instrumentation and Measurement

faster speed [67], a specific signature (features) of pneumonic disease [6, 38], and relatively economical [38, 40]. On the contrary, CT offers benefits such as better COVID region coverage of the lungs [63], stronger graphical feature and signature of pneumonic disease [9], bilateral change in COVID-19 infected patients with the ill-posed cases [49]. Finally, an effective tool in detection, quantification, and follow-up of the disease [19, 20]. Several studies [41, 51] have used both modalities to exploit the advantages of X-ray and CT. The percentage distribution for the use of image modalities were X-ray:45%, CT:50%, and X-ray+CT:5% is depicted in Figure 2 (b).

C.3 Statistical Distribution by Three Types of Pneumonia

We have categorized COVID-19 pneumonia into three categories based on the types of pneumonia. In binary (2-class) category, i.e., it is COVID *vs*. control, in ternary (3-class) category, i.e., COVID *vs*. control, *vs*. pneumonia, while in quaternary (4-class), i.e., COVID *vs*. control *vs*. viral pneumonia *vs*. bacterial pneumonia. The percentage distribution of these classes used in this study are depicted in Figure 2 (c) for *binary* (**62%**) [6, 8, 19, 20, 34, 35, 37, 42, 44, 47-49, 51, 52, 55-57, 59-64, 66, 67], for *ternary* (**28%)** [7, 9, 33, 36, 38, 41, 43, 46, 53, 58, 65] and for *quaternary* (**10%)** classification system [39, 40, 45, 50].

C.4 Statistical Distribution by Types of HDL Architectures

The architecture of HDL can broadly be classified into *spatial*, *temporal*, and *spatial-temporal,* as observed from various studies under HDL based on the classification of COVID images [68]. It was observed that only two studies [7, 41] (**5%**) used *spatial-temporal* HDL architecture, while 38 (**95%**) used *spatial* HDL architecture (see Figure 2 (d)).

C.5 Statistical Distribution by Data Size

Data size (DS) represents the number of images taken using the modalities of X-ray, CT, and both X-ray and CT. Since the data size drives HDL performance, it prevents over-fitting and imbalance. The distribution from 40 HDL studies is shown in Figure 3. (Note that only **17.5%** (7/40) of studies had dataset > 5K).

III. HDL Architecture, its Components, and Bias Estimation Strategy

A. Pipeline for ARDS diagnosis using HDL architecture

For a comprehensive RoB analysis, it is customary to investigate the basic building blocks of the ARDS pipeline in the HDL paradigm. As shown in Figure 4, the three major components of the ARDS pipeline are lung segmentation, COVID-19 severity classification, and lesion localization.

B. Typical HDL architecture and the Bias Concept

This section deals with the HDL architecture and the building blocks of HDL for COVID-19 diagnosis in the AI framework that leads to the motivation for bias estimation strategy. The HDL architecture is mainly driven by either the manifestation of three broad categories of imagery used (such as *spatial*, *temporal*, and *spatial-temporal*) [68] or the objective of the design of the HDL study for COVID-19 diagnoses, such as **LC**, **LC+LL**, **LS+LC**, and **LS+LC+LL** (section II.C). The studies considered under this systematic bias estimation (SBE) followed the *spatial* HDL architecture as they consider spatial input modalities such as X-ray and CT images while maintaining their corresponding *design objective* for COVID-19 detection.

A typical example of the HDL architecture is shown in Figure 5. Primarily, the HDL architecture considered CT scans and then does the separate branch of **LC+LL** followed by **LS**. Finally, the system performed a joint diagnosis of both branches for conformability. The other AI attributes contributing to the HDL architecture are the pre-processing, data augmentation, and high-value engineering performance which could provide the study's *low-bias* nature [30]. Furthermore, specific clinical objectives should be met, and if not met, they can cause a *high-bias* effect. If this HDL engineering attributes over-perform against the desired clinical outcomes, the HDL system is biased. Note that these AI attributes are responsible for the HDL performance that can be quantified against its optimal value. The difference in HDL's ability to reach optimal value against clinical performance is categorized as bias. Thus, we must have methods to represent the bias graphically or pictorially in the form of a map. Thus, the spirit of our innovation is the design of a map using a *radial* strategy that represents the fundamental core of HDL, i.e., the AI attributes. The quantification of such method leads to two more solutions, such as *regional-bias* area (RBA) and *ranking-bias* score (RBS) to estimate bias, thus evolving three such innovations to be discussed in the next section.

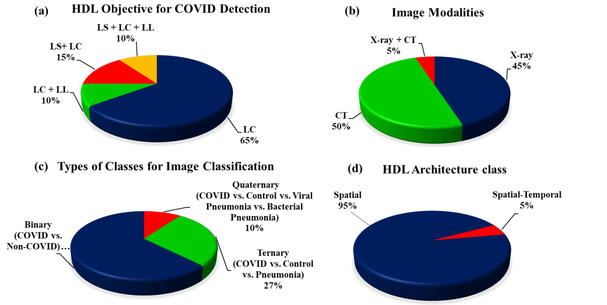


Figure 2. Statistical distribution by various criteria: (a) by objective; (b) by image modalities (c) by types of Pneumonia classes, and (d) by type of HDL architectures.

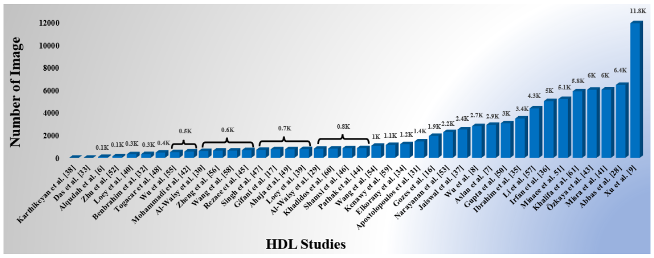


Figure 3. The distribution of increasing DS in various HDL studies for COVID-19 diagnosis. K~1000.

IV. Three Novel Paradigm

The concept of the RBM originates from the idea that when a system is unbalanced, there is always a leak, and this leak bleeds and spreads in a unique direction causing a protrusion. The strength of the leak can be noticed when compared to the un-leaked region.

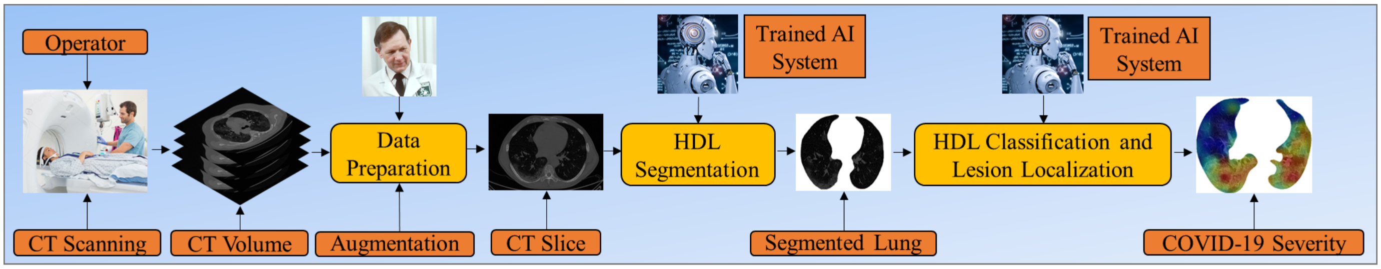


Figure 4. CT-based COVID diagnosis in ARDS using HDL.

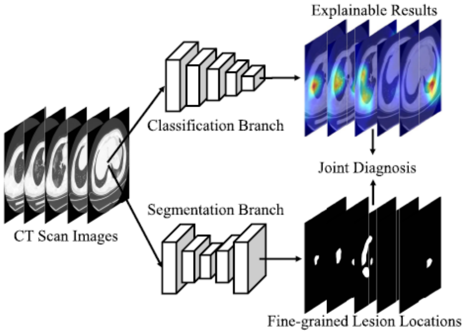


Figure 5. The standard architecture of HDL for COVID-19 diagnosis with low-bias effect [69] (permission pending).

The proposed system consists of AI and its attributes, and these attributes when made to spread out in 360 directions, create a map. When some of the attributes are too strong, while others are weak, it causes a dent in the system leading to a glitch, so-called bias. We have categorized this system as a "*radial-bias"* map (RBM) since the AI attributes are spanned in radial 360 directions. When some of the attributes are too strong, while others are weak, it causes a dent in the system leading to a glitch, so-called bias. We have categorized this system as a "*radial-bias"* map (RBM) since the AI attributes are spanned in radial 360 directions. Further, the map area between the maximum strength of the AI attributes that the system (study) can possess, and the minimum strengths of the AI attributes can also represent an indirect measure of the bias, called the *regional-bias* area (RBA). In the 3rd bias category, we adapt the score for each attribute and for each study leading to a cumulative score per study. These scores are then ranked, and bias cut-off is estimated. The process of bias identification using such a ranking paradigm is the "*rank-bias"* score (RBS) method.

A. Innovation-I: Radial-Bias Map

Since the HDL technology applied for COVID diagnosis consists of different stages such as design, optimization, performance evaluation, and clinical application, one must look for the strengths of the AI attributes (A1 to A39 in Table A1, Supporting document) in these stages (so-called clusters). We observed that the distribution of AI attributes in each of the four clusters were 14, 7, 8, and 10, respectively.

For estimating the strengths of AI attributes, we used a pictorial representation of the "spokes and wheel model" in 360 directions, where each spoke represents the product of the weight of the attribute times the radius of the spoke. The *bias value* (β_{radial}) measurement pseudo algorithm is summarized as follows: (**i**) Divide the AI attributes into four clusters (design, optimization, performance evaluation, and clinical validation) based on the HDL pipeline. (**ii**) Compute the spoke length of each AI attribute (weight x 80% of half the image size (256)). (**iii**) Compute the sum of spoke lengths corresponding to four clusters (say Σ_{C1}, Σ_{C2}, Σ_{C3}, and Σ_{C4}). (**iv**) Compute the sum of the top two and bottom two clusters (say Σ_A and Σ_B). (**v**) Compute the $\boldsymbol{\beta}$radial = $|\Sigma_A$-$\Sigma_B|$, as the absolute difference between Σ_A and Σ_B. (**vi**) The normalized bias value ($\boldsymbol{\beta}_{\text{radial}}^{\text{norm}}$)=($\frac{\boldsymbol{\beta}\text{radial}}{\alpha}$), where α is the total number of AI attributes. The weight matrix presents the weights of the AI attributes based on the experience and judgment of AI professionals. In all, each study has 39 attributes corresponding to every 9.2 (~360/39) degrees. The Bezier spline curve is then fitted through the endpoint of each spoke to represent the smooth curve. Since the curve has four sectors (corresponding to four clusters), the *radial-bias* map resembles butterfly wings, as shown in Figure 6, laid out in a 5x8 grid, representing 40 HDL studies. These studies are arranged from *low* to *high-bias*, where the bias of each study is in the corner of the *radial-bias* map (where the name of the bias map is: "Sn-Name:BiasValue", for example, "S31-Asl:18", where "31" represents the study number, "Asl" is the first three letters of the last name of the first author in the study, and "18" represents the normalized value of the bias). Note that the following is the sequence of AI attributes for each of the four clusters (A1 to A39 in Table A1, Supporting document). The AI *design cluster* (A1-A14) consisted of (i) HDL class, (ii) generalized hybrid (coarse), (iii) specialized hybrid (refined), (iv) a number of solo deep learning (SDL) architecture used to form HDL, (v) number of classifiers, (vi) several classes for the classification system, (vii) feature extraction methodology, (viii) feature selection methodology, (ix) pre-processing, (x) data augmentation, (xi) data partition scheme, (xii) a number of performance evaluation parameters for evaluation, (xiii) hardware and (xiv) software resource used. The *second cluster* (A15-A21) of AI-based attributes are the *seven optimization parameters* used in the HDL study. These are the (i) regularization method adapted, (ii) number of regularization methods, (iii) optimizers, (iv) loss function, (v) learning rate, (vi) batch size, and (vii) epochs of the HDL system. The *third cluster* (A22-A29) of attributes includes the *performance evaluation parameters* such as (i) accuracy, (ii) sensitivity, (iii) specificity, (iv) precision, (v) F1-score, (vi) Kappa, (vii) area-under-the-curve (AUC), and (viii) statistical analysis. The *last and fourth cluster* (A30-A39) consists of ten *benchmarking and clinical validation parameters* attributes. These include the (i) benchmarking with the number of models, (ii) clinical validation, (iii) scientific validation, (iv) image modality, (v) dataset size, (vi) performance analysis metrics, (vii) study objective, (viii) demographic data of dataset, (ix) clinical validation of dataset by a radiologist, and (x) RT-PCR test conducted to confirm the data on the dataset.

B. *Innovation-II: Regional-Bias Area*

The RBA is calculated by the area difference between the best AI performing attributes and the worst AI performing attributes. The RBA is depicted in Figure 7 for each study in increasing order of bias area, where the white region represents the bias area. Each study bias is represented as: "Sn-Name:BiasValue", for example, "S15-Rez:69", where "15" represents the study number, "Rez" is the first three letters of the last name of the first author in the study, and "69" represents the normalized value of the bias. Note, the more the bias area, the higher is the white shaded region.

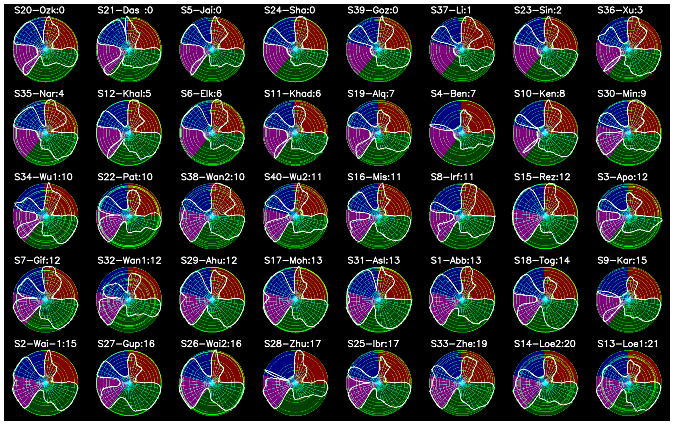


Figure 6. Demonstration of the results of *radial-bias* maps for 40 HDL studies in the order of the decreasing area of the spline-fitted butterfly. The studies are labeled as S1 to S40. The two-digit number after ":" is the normalized bias value of the *radial-bias* map.

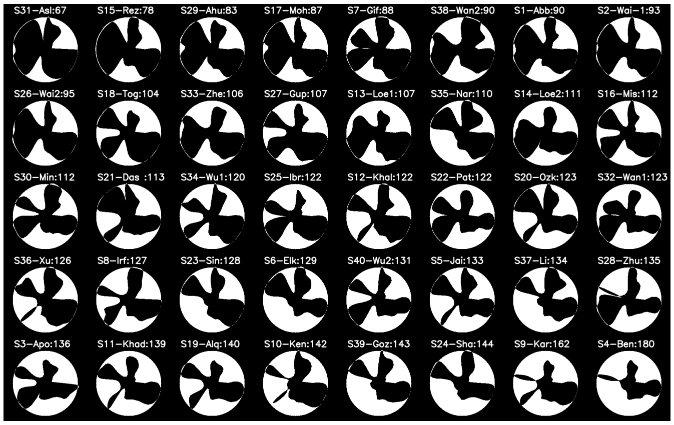


Figure 7. Bias configuration of HDL studies by *regional-bias* area method. White patches show the *regional-bias* area.

C. *Innovation-III: Ranking-Bias Score*

There are **40** HDL studies in AI under consideration for the COVID-19 diagnosis [6-9, 19, 20, 33-53, 55-67]. We created **39** AI-based attributes for each study, thus a total of 1,560 attributes. These HDL features are initially qualitative and then quantified by assigning a number between 0 and 5 based on the AI scientist's experience. The study's aggregate score is the sum of all attribute values for that selected study. Using the aforementioned technique, we plotted the mean values of the 40 HDL investigations, which ranged from 2.1 (right) to 4.0 (left), plotted in decreasing order (4.0 to 2.1), as shown in Figure 8 and Table A1 (Supporting document). We follow up this *ranking-bias* score method to find the bias in the HDL studies. The higher the mean value, the lower is RoB. Hence, the studies were arranged in the order of *low-bias*, *moderate-bias*, and *high-bias*, according to the decreasing order of their aggregate scores. The raw-cutoff of **2.9** was determined to select AI-based HDL studies for RoB based on the intersection of the "cumulative plot of the mean score and the frequency plot curve of the studies". According to the ranking score graph, the majority of the studies had a *moderate-bias* (ranging from 3.5 to 2.9, in decreasing order left to right, Figure 8), and this accounted for 24 studies (60%)). Note that all the moderate-bias studies were published simultaneously (in 2020) and did not offer more extensive diversity in the AI techniques for COVID diagnosis. There was a subtle change in the AI attributes between the studies. Note that the studies with higher normalized mean values in the AI attributes were considered as *low-bias*. These low-bias studies [7, 20, 33, 47, 50, 57, 62] showed more innovation in the design for COVID diagnosis. On the contrary, the tail-enders [6, 19, 36, 37, 43, 51, 54, 55, 65] showed low AI attribute mean scores (*high-bias*) and were not clinically substantial compared to *low-bias* or *moderate-bias* studies. We will discuss the analysis of the studies between the three quantitative and innovation methods in the next section.

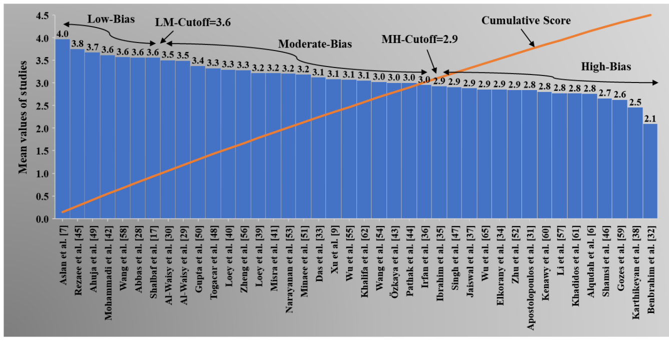


Figure. 8. Results of the *ranking-bias* score method showing the frequency distribution of HDL studies in decreasing order followed by the cumulative plot, showing the raw HDL cut-off. **LM**: Low-moderate cutoff 3.6, **MH**: Moderate-high cutoff 2.9.

V. Non-Randomized AI Methods and Inter-Comparison of Five Bias Methods

Recently two qualitative methods (ROBINS-I [31] and PROBAST [32]) were designed for non-randomized AI trials. The section converts the HDL's qualitative to quantitative measure using ROBINS-I and PROBAST paradigms.

A. *ROBINS-I*

The goal of this bias estimation method is to simulate the randomization of non-randomized trials. In order to study RoB, it covers seven different features (domains) that are grouped into three intervention components (marked parameters): (a) "Pre-Intervention," (b) "During Intervention," and (c) "Post-Intervention." Table A2 (Supporting document) illustrates as: **(C1)** *confounding factors* (data size, data source, and inclusion of demographic data), **(C2)** *participant selection* (partitioning of dataset and number of HDL models included), **(C3)** *intervention classifications* (imaging features, pre-processing, and data augmentation), **(C4)** *intended deviation* (validation and verification by the radiologist and benchmarking), **(C5)** *missing data* (SWAB test and usage of loss function, optimizers, and regularization), **(C6)** the *measurement of outcome* (prevents it from being included in the meta-analysis, performance evaluation parameters), and **(C7)** *result reporting* (clinical validation and statistical analysis). In order to represent the results of the qualitative analysis, a three-color scheme was used in the HDL framework. The *red* signifies a *high-bias* in the study, indicating a serious problem with the AI attributes taken into consideration, and the attribute was given a score of 1. Those with a *moderate-bias* (*yellow)* were given a score of 3, whereas those with a *low-bias* (*green)* performed well compared to the testing parameters and were given a score of 5. These scores were summed up, and the final mean score was calculated for all the 40 studies. Using ROBINS-I, **~55%** (22 out of 40) studies had *high-bias*, and 14 studies (~**35%**) were *moderate-bias* (Figure 9 c-d). There were four studies in the *low-bias* zone (Figure 9 b).

B. PROBAST

The PROBAST is a prominent RoB assessment tool based on AI with the following features (a) *participants*, whether or not a radiologist validated them, type of data source, and demographics; (b) *predictors*, consisting of imaging features, pre-processing, data augmentation, loss functions, and optimizers; (c) *outcomes*, if RT-PCR test was performed for the cohort, and performance evaluation parameters were evaluated and (d) *analysis,* if it would cover the data partitioning, patient count, benchmarking against other models, clinical validation, and statistical evaluation. The ranking was performed using **AP(ai)Bias 2.0**, (Table A3, Supporting document), using the same 40 studies. With the use of PROBAST, we found that ~**35%** percent (14 out of 40) of the studies showed a *high-bias* (red), ~**45%** (18 out of 40) were *moderately-biased*, and eight studies were in the *low-biased* (Figure 9 b-d).

C. Analysis of Three Bias Strategies: Venn diagram

This section represents the Venn diagram (VD) approach to analyze the relationship between the three innovative methods (RBM *vs.* RBA *vs.* RBS) for RoB. Figure 9 (a) depicts the process of the VD under three categories of bias such as (a) *low-bias*, (b) *moderate-bias*, and (c) *high-bias*. The number of studies in *low-bias* for RBM, RBA, and RBS were **16** (40%) [6-8, 33, 35, 36, 42, 46-50, 55, 57, 59, 62], **11** (27.5%) [7, 20, 33-35, 47, 50, 53, 57, 64, 66] , and **7** (17.5%) [7, 20, 33, 47, 50, 57, 62] respectively. The number of studies under *moderate-bias* for RBM, RBA, and RBS were **9** (22.5%) [20, 34, 40, 53, 54, 58, 60, 63, 66], **11** (27.5%) [38, 40, 44-46, 49, 56, 58, 59, 61, 63], and **24** (60%) [8, 9, 34, 35, 38-42, 44-46, 48, 49, 52, 53, 56, 58-64], respectively. Similarly, for *high-bias* were **15** (37.5%) [9, 19, 37, 38, 41, 43-45, 51, 52, 54, 56, 61, 64, 65], **18** (45%) [6, 9, 19, 36, 37, 39, 41-43, 48, 51, 52, 54, 55, 60, 62, 63, 65], and **9** (22.5%) [6, 19, 36, 37, 43, 51, 54, 55, 65]. The studies that fall under the intersection of *low-bias*, *moderate-bias*, and *high-bias* were **5**, **3**, and **6**, respectively for the three innovative methods (Figure 9 (a)).

D. Analysis of Bias using five measures

To create a VD, the following steps were used (Figure 9 b-d). ROBINS-I and PROBAST were converted from qualitative to quantitative schemes using conversion scores such as 5 for *low-bias*, 3 for *moderate-bias*, 1 for high-bias, and 0 for *unclear-bias*. (ii) Common studies are shown between (a) *radial-bias* map, (b) *regional-bias* area, (c) *ranking-bias* score (d) ROBINS-I, and (e) PROBAST. (iii) The same set of 40 HDL studies was normalized into digital count, as shown in Figure 9 (a-d).

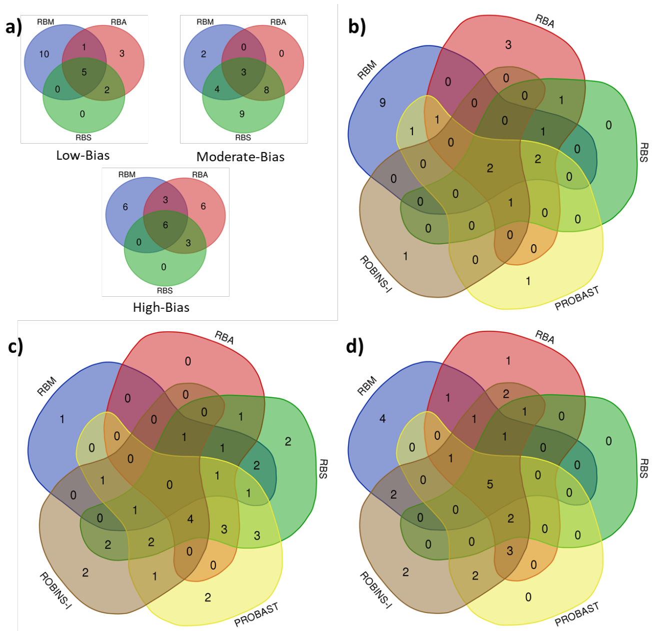


Figure 9. Comparison of (a) three analytical bias methods (RBM, RBA, and RBS) using VD. (b), (c), and (d) shows the comparison of five bias methods (RBM, RBA, RBS, PROBAST, and ROBINS-I) low-bias, moderate-bias, and high-bias measurements, respectively.

E. Cluster analysis between two groups: three quantitative methods (Gr. A) and two non-randomized AI methods (Gr. B)

Given two clusters (say group A & group B, where A is a pool of three innovative methods: RBM, RBA, and RBS methods while B is a pool consisting of two older methods: ROBINS-I and PROBAST). It is important to investigate the intersection of the two clusters in *moderate* and *high-bias* categories. The intersection is defined as a combination of 2, 3, 4, and 5 bias methods using clusters A & B. This is shown in Figure 10, where yellow and red represents *moderate-bias* and *high-bias*, respectively. Following are the conclusions (i) High-bias studies strongly overlap between the two clusters. This is because most of the studies do not do (a) clinical validation, (b) the datasets were not verified and validated by the radiologist, (c) feature selection was not performed, and (d) the RT-PCR test was not conducted. (ii) *Moderate-bias* has a lower frequency compared to the *high-bias*. The mean frequency for *moderate-bias* and *high-bias* studies were **3.19** and **8.38,** respectively. (iii) Overlap between the clusters was ~**163%** ((3.19-8.38) / 3.19) more in the *high-bias* compared to the *moderate-bias*. Figure 11 shows that with the decrease in the cutoff from 4 to 2.1, more studies participate in the *low-bias* region. Only one study passed the acceptability criteria of 80%

[7], and with the cutoff of 3.6, seven studies passed the hypothesis.

VI. Discussion

This is the first study of its kind to demonstrate three new methods for AI bias estimation in the HDL paradigm while considering ARDS under the class of **AP(ai)Bias 2.0**. Forty studies were selected using the PRISMA model, a well-established standard in the healthcare industry. The study showed various statistical distributions by various criteria such as (a) by objective; (b) by image modalities (c) by types of Pneumonia classes, and (d) by type of HDL architectures. Note that HDL is not same as ensemble or Hybrid Ensemble Deep Learning Model (HEDL). This is because ensemble is a combination of several classification methods, while HEDL is combination of ensemble and hybrid [70]. The main novelty of our study was the analytical design of three methods RBM, RBA, RBS, and subsequently validated against two of the previously developed non-randomized AI strategies such as ROBINS-I, and PROBAST. The RBM was the most elegant method since it demonstrated the map of high-performing vs. low-performing AI attributes in a study. To this complement, the RBA method offers a regional area method for joint visualization and bias computation. The RBS method used the aggregate score paradigm followed by ranking in the HDL framework. We analyzed these systems using the Venn diagram. Finally, based on moderate-high and low-moderate cutoffs of **2.9** and **3.6**, respectively, we observed **40%**, **27.5%**, **17.5%**, **10%**, and **20%** studies were low-biased for RBM, RBA, RBS, ROBINS-I, and PROBAST, respectively. Our system **AP(ai)Bias 2.0** used 39 AI attributes on 40 HDL studies unlike the previous study that used 10 AI attributes on 42 DL studies under the class of **AP(ai)Bias 1.0**. Considering the two pools A and B, where pool A consisted of RBM, RBA, and RBS and pool B consisted of ROBINS-I and PROBAST, we showed the inter-combinations of clusters between the two pools. There was 163% more overlap between pool A designed using **AP(ai)Bias 2.0** and pool B (previous methods) in *high-bias* compared to *moderate-bias*.

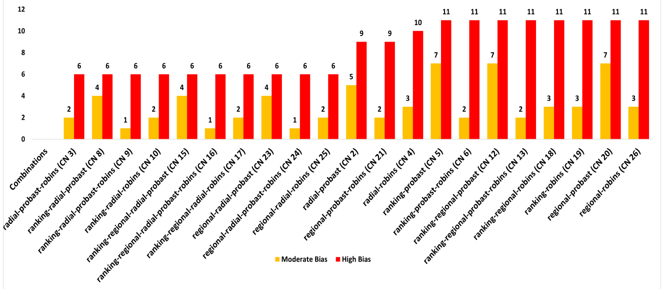


Figure 10. Inter-combinations of clusters between the two pools (A is a pool of RBM, RBA, and RBS while pool B is ROBINS-I and PROBAST). Yellow: *moderate-bias* and Red: *high-bias*.

A short note on over-emphasis on classification paradigm

The robustness of any HDL model for COVID-19 detection can be reflected by its performance evaluation (PE) parameters. The standard PE parameters used by the HDL-COVID-19 detection system for classification are accuracy, sensitivity, specificity, precision, F-1 score, Kappa, and area under the curve (AUC). However, accuracy is a very well-known parameter and is adapted by almost all HDL models.

Benchmarking Table

The benchmarking Table 1 shows a comparison between our proposed work with *six* other studies [28, 71-75], where **13** attributes were considered. The proposed study is in the last column. Note that we offer "✓" in places for unique contribution in the proposed model and "×"in the absence of any other contribution.

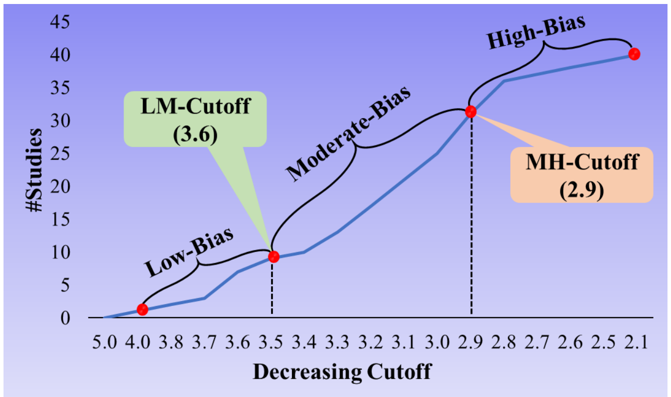


Figure 11. Plot showing the number of studies in the *low*, *moderate*, and *high-bias* regions with the decreasing cutoff.

Table 1. Benchmarking table.

SN	Attributes	Alzahab *et al.* [71]	Kao *et al.* [74]	Albahari *et al.* [72]	Roberts *et al.* [75]	Bao *et al.* [73]	Suri *et al.* [28]	Suri *et al.* (Proposed)
1	Date	Jan. 2021	May 2021	June 2020	Oct. 2020	Jun. 2020	Aug. 2021	2021
2	AI Spec.	HDL	AI	AI	ML	AI	DL	HDL
3	Application	BCI	COVID	COVID	COVID	COVID	COVID	COVID
4	RBM	×	×	×	×	×	×	✓
5	RBA	×	×	×	×	×	×	✓
6	RBS	×	×	×	×	×	✓	✓
7	ROBINS-I	×	×	×	×	×	✓	✓
8	PROBAST	×	×	×	✓	×	✓	✓
9	Other	×	Funnel	×	×	×	×	×
10	PRISMA	✓	✓	✓	✓	✓	✓	✓
11	# of study	47	6	11	45	13	89	42
12	References	96	49	109	84	30	116	67
13	Objective	LS, LC	LC	LC	LC	LC	LC	LS, LC, LL

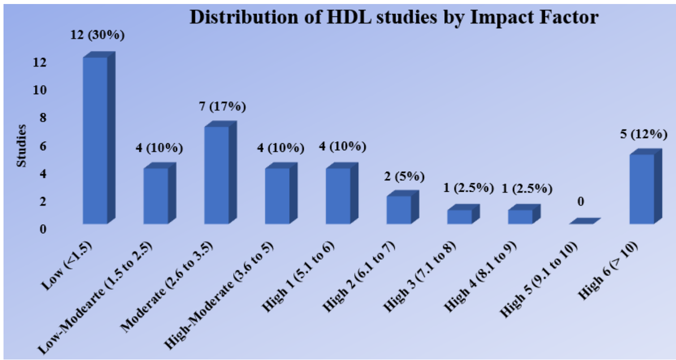


Figure 12. Bias analysis and Impact factor of studies.

Comparison between Suri et al. (JBHI'21) & **AP(ai)Bias 2.0**

The most fundamental difference between the current study and Suri *et al.* [28] is the design of three novel bias methods (such as RBM, RBA, RBS, so-called **AP(ai)Bias 2.0**) while considering 40 HDL studies using 39 AI attributes, unlike in Suri *et al.* [28] (JBHI '2021) considering 42 DL studies using 10 AI attributes, so-called **AP(ai)Bias 1.0.** The underlying principle in **AP(ai)Bias 2.0,** is the vector representation of the AI attribute. It is distributed circularly in four clusters leading to butterfly wings, obtained using the "spokes and wheel" model. **AP(ai)Bias 2.0** was fully automated by accurately computing the RBM, RBA, and RBS measurements. Finally, the low-bias intersection between the five bias methods is elegantly presented, unlike in **AP(ai)Bias 1.0**, where the intersection is of 3 methods only. Note that the cutoffs in **AP(ai)Bias 1.0** was 1.9, unlike in **AP(ai)Bias 2.0**, the cutoff was 3.6. In **AP(ai)Bias 2.0**, for low bias, a raw cutoff of 3.6 was computed using RBS. Using the cutoff of 3.6, it was discovered that RBM, RBA, RBS, ROBINS-I, and PROBAST had only 40%, 27.5%, 17.5%, 10%, and 20% studies in the *low-bias*, respectively. Only one study qualified in *low-bias* category. On the contrary, **AP(ai)Bias 1.0**, showed that ROBINS-I and PROBAST had only 32%, 16%, and 26% studies, respectively in *low-moderate* RoB (cutoff>2.5), and none of them qualified for the RoB hypothesis. Overall, the standard of HDL studies was better having a category of *moderate-bias,* unlike in the DL framework where most of the studies were *high-biased*.

A short note on Bias analysis and Impact Factor of Studies

We uncovered why the bias showed higher value in these 40 HDL studies. Quality has always been a factor in publication. Figure 12 shows the impact factor (IF) distribution of 40 HDL studies for COVID-19 diagnosis. Keeping our IF threshold of 5, only 12% of the studies were published in IF>6.0. The standard to see was that 60% of the studies were published in IF<3.5, while 30% were published in IF<1.5. Thus, this is one reason for *high-bias* in studies since the objective was to quickly publish fast in journals with low IF. Further, one reason which accounts for *high-bias* is the lack of thorough research. This directly points to the cost of conducting research, funding for research, multicenter data access, and lack of participation by radiology centers (which could be due to their interest and no direct incentive for the participating radiologists).

Recommendations/Challenges

The proposed study presents several recommendations that can improve the AI-bias in forthcoming studies. We have clustered these recommendations based on the stages of the pipeline such as (a) objective-clarity, data size and ground truth clinical information, (b) design of the HDL architecture and the optimization parameters, (c) performance evaluation of the engineering parameters, (d) scientific and clinical validation, hardware constraints, and finally the resources which includes funding for the entire project. Overall, we have eight-point crucial recommendations discussed below: **(i)** *Objective and vision of the study:* This should be clearly defined based on the four types of objectives. This can include (i) **LC** for COVID-19 detection; (ii) **LC+LL**; (iii) **LS+LC**; or (iv) **LS+LC+LL**. This should consider if the system being developed is *spatial*, temporal, or *spatial-temporal*. Lastly, it should take into consideration the type of classification such as *binary* (two classes), *trinary* (three classes), and *quaternary* (four classes). Finally, dimensionality (2D *vs*. 3D) should be taken into consideration when choosing the vision. **(ii)** *Data Size and Ground Truth Clinical Information*: Only **50**% of the studies had #CT/X-ray scans >3K and **17.5**% studies >5K. This becomes challenging if the objective classification of multiclass (classes >3-10). Balancing and augmentation is one solution, but this introduces bias. Thus, to avoid AI-bias, it is required to have a multicenter data collection of data sizes >10K. As part of the ground truth information, one must collect clinical information such as ground-glass opacities (GGO), type of pneumonia (such as COVID, bacterial, viral-community, atypical, influenza, and legionnaire), grading of the COVID-19 severity (*low*, *moderate*, and *high*), location and annotation of the lesions inside the lung region. **(iii)** *HDL architectural vision*: HDL architecture should consider how the data is trained using an AI model. Since the COVID-19 disease has different repercussions in patients having comorbidity such as renal disease, coronary artery disease, neurological disease, diabetes, peripheral disease, etc., the training models can be designed based on "COVID-19 lung CT scans severity along with its comorbidity". This way, the appropriate model is not applied to the "unseen CT scans" with specific symptoms. Thus, HDL training models should be tied to comorbidity to avoid the AI-bias and to bleed in the *radial-bias* map. This will lead to the best design with the least AI-bias. Further, the need for robust initial weights during transfer learning must be used to avoid re-training of the deep learning systems. **(iv)** *Optimization of HDL Architecture*: Optimization in engineering design must be conducted for best HDL architecture, such as the type of *optimizer* (ADAM, root mean squares (RMSprop), stochastic gradient descent (SGD)), type of *loss functions* (Cross-Entropy, Dice Similarity Coefficient, Hinge, Empirical), epochs needed, batch normalization, depth of the neural network, and learning rates. **(v)** *Performance Evaluation Parameters*: The performance evaluation of the system design must be conducted for all AI attributes (360) which should give equal importance to scientific parameters (classifier parameter evaluations) and clinical parameters (statistical parameter evaluations). These performances must have a feedback loop to overcome the weak parameters to avoid AI-bias. ***(vi)*** *Scientific Validation and Clinical Evaluation*: The outcome of the AI study on COVID-19 data must be scientifically and clinically validated. The current gold standard is the RT-PCT test and must be conducted as part of the study. Scientifically, the system must be validated on unseen data that is not part of the training system. Verification of the software must be conducted to avoid failure in the software design. ***(vii)*** *Hardware and Software Requirement*: AI training models seldom require large data size (a) having 512x512 to 1024x1024 sized images, (b) 2-3 bytes per pixel (8-16-24 bits per pixel), and (c) big cohorts. This all leads to large memory and processing power requirements. Thus, GPU or GPU clusters are sometimes needed to avoid cutting corners and AI-bias. ***(viii)*** *Funding*: At national and international levels, collaborations must be conducted at the United Nations Organization (UNO) level to create funds for research groups to expedite the dedicated scientists who cannot access funds in their own countries. This should be non-political and solely based on scientific merits providing inputs to designs that can prevent AI-bias in outcomes.

Strength, Weakness, and its Extension

AP(ai)Bias 2.0 collectively offers three innovative solutions for AI-bias estimation using RBM, RBA, and RBS. Further, the **AP(ai)Bias 2.0** was benchmarked against two non-randomized AI-bias estimation methods (ROBINS-I and PROBAST) by converting qualitative measures into **AP(ai)Bias 2.0** framework. The main strength of **AP(ai)Bias 2.0** was its fully automated design using Python language offering spline fitted butterfly maps, RBM, and RBS along with its cutoff in a click of the button, given the weight matrix. The major weakness of the system was the lack of information in the published studies which was considered as low-weights or score. While this system is truly innovative, fast, reliable, and designed with experienced team consent, variability studies and fusion ensemble methods need to be conducted for further validation [76, 77]. Search criteria need to broaden by including keywords such as “fusion, combine, cascaded AI models” [78]. The application of the bias methods is not restricted to COVID-19 ARDS application alone and can be extended to other applications such as cardiovascular risk stratification [79, 80], brain tumor [81], and Parkinson's disease [82] using AI paradigms.

VII. Conclusion

This is the first study on COVID-19 diagnosis using HDL that envelops three innovative and powerful solutions for bias estimation in AI by using **AP(ai)Bias 2.0** which consists of a *radial-bias* map, *regional-bias* area, and *ranking-bias* score. **AP(ai)Bias 2.0** was benchmarked against ROBINS-I and PROBAST, demonstrating consistent results for the three bias bins (*low*, *moderate*, and *high*). The bias was analyzed using a Venn diagram between (a) three innovative methods and (b) among the five RoB models. Based on the cumulative score of the ranking paradigm having a cutoff of **3.6**, the percentage of low-bias studies in the five pools were **40%**, **27.5%**, **17.5%**, **10%**, and **20%**, corresponding to RBM, RBA, RBS, ROBINS-I, and PROBAST, respectively. Finally, the study HDL presented a set of eight-point recommendations for minimizing the AI-bias.

References

[1] "WHO Coronavirus (COVID-19) Dashboard." [Online]. Available: https://covid19.who.int/.

[2] Johns Hopkins University (JHU) covid 19 dashboard [Online]. Available: https://coronavirus.jhu.edu/map.html

[3] A. C. Darby and J. A. Hiscox, "Covid-19: variants and vaccination," ed: British Medical Journal Publishing Group, 2021.

[4] D. John, J. Menon, G. Jammy, and A. Banerjee, "Estimation of the economic burden of COVID-19 using Disability-Adjusted Life Years (DALYs) and Productivity Losses in Kerala, India," *BMJ open,* vol. 11, no. 8, 2021.

[5] S. L. Emery *et al.*, "Real-time reverse transcription-polymerase chain reaction assay for SARS-associated coronavirus," *Emerg Infect Dis,* vol. 10, no. 2, pp. 311-6, Feb 2004, doi: 10.3201/eid1002.030759.

[6] A. M. Alqudah, S. Qazan, H. Alquran, I. A. Qasmieh, and A. Alqudah, "Covid-2019 detection using x-ray images and artificial intelligence hybrid systems," vol. 2, no. 16077.59362, p. 1, 2020, doi: https://doi.org/10.13140/RG.

[7] M. F. Aslan, M. F. Unlersen, K. Sabanci, and A. Durdu, "CNN-based transfer learning-BiLSTM network: A novel approach for COVID-19 infection detection," *Appl Soft Comput,* vol. 98, p. 106912, Jan 2021, doi: 10.1016/j.asoc.2020.106912.

[8] Y. H. Wu *et al.*, "JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation," *IEEE Trans Image Process,* vol. 30, pp. 3113-3126, 2021, doi: 10.1109/TIP.2021.3058783.

[9] X. Xu *et al.*, "A Deep Learning System to Screen Novel Coronavirus Disease 2019 Pneumonia," *Engineering (Beijing),* vol. 6, no. 10, pp. 1122-1129, Oct 2020, doi: 10.1016/j.eng.2020.04.010.

[10] U. R. Acharya *et al.*, "An accurate and generalized approach to plaque characterization in 346 carotid ultrasound scans," *IEEE transactions on instrumentation and measurement,* vol. 61, no. 4, pp. 1045-1053, 2011.

[11] U. R. Acharya *et al.*, "Plaque tissue characterization and classification in ultrasound carotid scans: a paradigm for vascular feature amalgamation," *IEEE Transactions on Instrumentation and Measurement,* vol. 62, no. 2, pp. 392-400, 2012.

[12] S. Delsanto, F. Molinari, P. Giustetto, W. Liboni, S. Badalamenti, and J. S. Suri, "Characterization of a completely user-independent algorithm for carotid artery segmentation in 2-D ultrasound images," *IEEE Transactions on Instrumentation and Measurement,* vol. 56, no. 4, pp. 1265-1274, 2007.

[13] R. Cau, P. P. Bassareo, L. Mannelli, J. S. Suri, and L. Saba, "Imaging in COVID-19-related myocardial injury," *Int J Cardiovasc Imaging,* vol. 37, no. 4, pp. 1349-1360, Apr 2021, doi: 10.1007/s10554-020-02089-9.

[14] J. S. Suri *et al.*, "COVID-19 pathways for brain and heart injury in comorbidity patients: A role of medical imaging and artificial intelligence-based COVID severity classification: A review," *Comput Biol Med,* vol. 124, p. 103960, Sep 2020, doi: 10.1016/j.compbiomed.2020.103960.

[15] R. Cau *et al.*, "Complications in COVID-19 patients: Characteristics of pulmonary embolism," *Clin Imaging,* vol. 77, pp. 244-249, Sep 2021, doi: 10.1016/j.clinimag.2021.05.016.

[16] L. Saba *et al.*, "Molecular pathways triggered by COVID-19 in different organs: ACE2 receptor-expressing cells under attack? A review," *Eur Rev Med Pharmacol Sci,* vol. 24, no. 23, pp. 12609-12622, Dec 2020, doi: 10.26355/eurrev_202012_24058.

[17] V. Viswanathan *et al.*, "Bidirectional link between diabetes mellitus and coronavirus disease 2019 leading to cardiovascular disease: A narrative review," *World journal of diabetes,* vol. 12, no. 3, p. 215, 2021.

[18] D. Fanni *et al.*, "Vaccine-induced severe thrombotic thrombocytopenia following COVID-19 vaccination: a report of an autoptic case and review of the literature," *Eur Rev Med Pharmacol Sci,* vol. 25, no. 15, pp. 5063-5069, Aug 2021, doi: 10.26355/eurrev_202108_26464.

[19] O. Gozes *et al.*, "Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis," *arXiv preprint arXiv:.05037,* 2020.

[20] A. Shalbaf and M. Vafaeezadeh, "Automated detection of COVID-19 using ensemble of transfer learning with deep convolutional neural network based on CT scans," *International journal of computer assisted radiology surgery,* vol. 16, no. 1, pp. 115-123, 2021.

[21] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, "COVID-CT-dataset: a CT scan dataset about COVID-19," *arXiv preprint arXiv:.13865,* 2020.

[22] R. Cau *et al.*, "Computed tomography findings of COVID-19 pneumonia in Intensive Care Unit-patients," *J Public Health Res,* vol. 10, no. 3, Apr 19 2021, doi: 10.4081/jphr.2021.2270.

[23] L. Saba *et al.*, "The present and future of deep learning in radiology," *Eur J Radiol,* vol. 114, pp. 14-24, May 2019, doi: 10.1016/j.ejrad.2019.02.038.

[24] A. El-Baz and J. S. Suri, *Big Data in Multimodal Medical Imaging*. Boca Raton: CRC Press, 2019.

[25] L. Saba *et al.*, "Six artificial intelligence paradigms for tissue characterisation and classification of non-COVID-19 pneumonia against COVID-19 pneumonia in computed tomography lungs," *Int J Comput Assist Radiol Surg,* vol. 16, no. 3, pp. 423-434, Mar 2021, doi: 10.1007/s11548-021-02317-0.

[26] M. Agarwal *et al.*, "A novel block imaging technique using nine artificial intelligence models for COVID-19 disease classification, characterization and severity measurement in lung computed tomography scans on an Italian cohort," *Journal of Medical Systems,* vol. 45, no. 3, pp. 1-30, 2021.

[27] J. S. Suri *et al.*, "Integration of cardiovascular risk assessment with COVID-19 using artificial intelligence," *Reviews in Cardiovascular Medicine,* vol. 21, no. 4, pp. 541-560, 2020.

[28] J. S. Suri *et al.*, "Systematic Review of Artificial Intelligence in Acute Respiratory Distress Syndrome for COVID-19 Lung Patients: A

Biomedical Imaging Perspective," *IEEE J Biomed Health Inform,* vol. 25, no. 11, pp. 4128-4139, Nov 2021, doi: 10.1109/JBHI.2021.3103839.

[29] S. S. Sanagala *et al.*, "Ten Fast Transfer Learning Models for Carotid Ultrasound Plaque Tissue Characterization in Augmentation Framework Embedded with Heatmaps for Stroke Risk Stratification," *Diagnostics,* vol. 11, no. 11, p. 2109, 2021.

[30] P. K. Jain, N. Sharma, A. A. Giannopoulos, L. Saba, A. Nicolaides, and J. S. Suri, "Hybrid deep learning segmentation models for atherosclerotic plaque in internal carotid artery B-mode ultrasound," *Comput Biol Med,* vol. 136, p. 104721, Sep 2021, doi: 10.1016/j.compbiomed.2021.104721.

[31] J. A. Sterne *et al.*, "ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions," *BMJ,* vol. 355, p. i4919, Oct 12 2016, doi: 10.1136/bmj.i4919.

[32] R. F. Wolff *et al.*, "PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies," *Ann Intern Med,* vol. 170, no. 1, pp. 51-58, Jan 1 2019, doi: 10.7326/M18-1376.

[33] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "4S-DT: Self-Supervised Super Sample Decomposition for Transfer Learning With Application to COVID-19 Detection," *IEEE Transactions on Neural Networks Learning Systems,* vol. 32, no. 7, pp. 2798 - 2808, 2021.

[34] A. Al-Waisy *et al.*, "Covid-deepnet: hybrid multimodal deep learning system for improving covid-19 pneumonia detection in chest x-ray images," *Computers, Materials Continua,* vol. 67, no. 2, 2021.

[35] A. S. Al-Waisy *et al.*, "COVID-CheXNet: hybrid deep learning framework for identifying COVID-19 virus in chest X-rays images," *Soft comput,* pp. 1-16, Nov 21 2020, doi: 10.1007/s00500-020-05424-3.

[36] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Phys Eng Sci Med,* vol. 43, no. 2, pp. 635-640, Jun 2020, doi: 10.1007/s13246-020-00865-4.

[37] H. Benbrahim, H. Hachimi, and A. Amine, "Deep transfer learning with apache spark to detect covid-19 in chest x-ray images," *Romanian Journal of Information Science Technology,* vol. 23, no. S, SI, pp. S117-S129, 2020.

[38] N. N. Das, N. Kumar, M. Kaur, V. Kumar, and D. Singh, "Automated deep transfer learning-based approach for detection of COVID-19 infection in chest X-rays," *Irbm,* 2020.

[39] A. S. Elkorany and Z. F. Elsharkawy, "COVIDetection-Net: A tailored COVID-19 detection from chest radiography images using deep learning," *Optik,* vol. 231, p. 166405, 2021.

[40] A. U. Ibrahim, M. Ozsoz, S. Serte, F. Al-Turjman, and P. S. Yakoi, "Pneumonia classification using deep learning from chest X-ray images during COVID-19," *Cognitive Computation,* pp. 1-13, 2021.

[41] M. Irfan *et al.*, "Role of Hybrid Deep Neural Networks (HDNNs), Computed Tomography, and Chest X-rays for the Detection of COVID-19," *Int J Environ Res Public Health,* vol. 18, no. 6, p. 3056, Mar 16 2021, doi: 10.3390/ijerph18063056.

[42] A. Jaiswal, N. Gianchandani, D. Singh, V. Kumar, and M. Kaur, "Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning," *Journal of Biomolecular Structure Dynamics,* pp. 1-8, 2020.

[43] D. Karthikeyan, A. S. Varde, and W. Wang, "Transfer learning for decision support in Covid-19 detection from a few images in big data," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020: IEEE, pp. 4873-4881.

[44] M. Loey, G. Manogaran, and N. E. M. Khalifa, "A deep transfer learning model with classical data augmentation and CGAN to detect COVID-19 from chest CT radiography digital images," *Neural Comput Appl,* pp. 1-13, Oct 26 2020, doi: 10.1007/s00521-020-05437-x.

[45] M. Loey, F. Smarandache, and N. E. M Khalifa, "Within the lack of chest COVID-19 X-ray dataset: a novel detection model based on GAN and deep transfer learning," *Symmetry,* vol. 12, no. 4, p. 651, 2020.

[46] S. Misra, S. Jeon, S. Lee, R. Managuli, I.-S. Jang, and C. Kim, "Multi-channel transfer learning of chest X-ray images for screening of COVID-19," *Electronics,* vol. 9, no. 9, p. 1388, 2020.

[47] R. Mohammadi, M. Salehi, H. Ghaffari, A. Rohani, and R. Reiazi, "Transfer learning-based automatic detection of coronavirus disease 2019 (COVID-19) from chest X-ray images," *Journal of Biomedical Physics Engineering,* vol. 10, no. 5, p. 559, 2020.

[48] U. Özkaya, Ş. Öztürk, and M. Barstugan, "Coronavirus (COVID-19) classification using deep features fusion and ranking technique," in *Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach*: Springer, 2020, pp. 281-295.

[49] Y. Pathak, P. K. Shukla, A. Tiwari, S. Stalin, S. Singh, and P. K. Shukla, "Deep Transfer Learning Based Classification Model for COVID-19 Disease," *Ing Rech Biomed,* May 20 2020, doi: 10.1016/j.irbm.2020.05.003.

[50] K. Rezaee, A. Badiei, and S. Meshgini, "A hybrid deep transfer learning based approach for COVID-19 classification in chest X-ray images," in *2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME)*, 2020: IEEE, pp. 234-241.

[51] A. Shamsi *et al.*, "An Uncertainty-Aware Transfer Learning-Based Framework for COVID-19 Diagnosis," *IEEE Trans Neural Netw Learn Syst,* vol. 32, no. 4, pp. 1408-1417, Apr 2021, doi: 10.1109/TNNLS.2021.3054306.

[52] M. Singh, S. Bansal, S. Ahuja, R. K. Dubey, B. K. Panigrahi, and N. Dey, "Transfer learning–based ensemble support vector machine model for automated COVID-19 detection using lung computerized tomography scan data," *Medical biological engineering and computing,* vol. 59, no. 4, pp. 825-839, 2021.

[53] M. Toğaçar, B. Ergen, and Z. Cömert, "COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches," *Computers in biology and medicine,* vol. 121, p. 103805, 2020.

[54] E. M. El-Kenawy, A. Ibrahim, S. Mirjalili, M. M. Eid, and S. E. Hussein, "Novel Feature Selection and Voting Classifier Algorithms for COVID-19 Classification in CT Images," *IEEE Access,* vol. 8, pp. 179317-179335, 2020, doi: 10.1109/ACCESS.2020.3028012.

[55] A. Khadidos, A. O. Khadidos, S. Kannan, Y. Natarajan, S. N. Mohanty, and G. Tsaramirsis, "Analysis of COVID-19 Infections on a CT Image Using DeepSense Model," *Front Public Health,* vol. 8, p. 599550, 2020, doi: 10.3389/fpubh.2020.599550.

[56] N. E. M. Khalifa, M. H. N. Taha, A. E. Hassanien, and S. Elghamrawy, "Detection of coronavirus (COVID-19) associated pneumonia based on generative adversarial networks and a fine-tuned deep transfer learning model using chest X-ray dataset," *arXiv preprint arXiv:.01184,* 2020.

[57] S. Ahuja, B. K. Panigrahi, N. Dey, V. Rajinikanth, and T. K. Gandhi, "Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices," *Applied Intelligence,* vol. 51, no. 1, pp. 571-585, 2021.

[58] A. Gupta, S. Gupta, and R. Katarya, "InstaCovNet-19: A deep learning classification model for the detection of COVID-19 patients using Chest X-ray," *Applied Soft Computing,* vol. 99, p. 106859, 2021.

[59] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. J. Soufi, "Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning," *Medical image analysis,* vol. 65, p. 101794, 2020.

[60] J. Zhu, B. Shen, A. Abbasi, M. Hoshmand-Kochi, H. Li, and T. Q. Duong, "Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs," *PLoS One,* vol. 15, no. 7, p. e0236621, 2020, doi: 10.1371/journal.pone.0236621.

[61] B. N. Narayanan, R. C. Hardie, V. Krishnaraja, C. Karam, and V. S. P. Davuluru, "Transfer-to-transfer learning approach for computer aided detection of COVID-19 in chest radiographs," *AI,* vol. 1, no. 4, pp. 539-557, 2020.

[62] S. Wang *et al.*, "A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)," *European radiology,* pp. 1-9, 2021.

[63] X. Wu *et al.*, "Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: A multicentre study," *Eur J Radiol,* vol. 128, p. 109041, Jul 2020, doi: 10.1016/j.ejrad.2020.109041.

[64] C. Zheng *et al.*, "Deep learning-based detection for COVID-19 from chest CT using weak label," *MedRxiv,* 2020.

[65] L. Li *et al.*, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology,* 2020.

[66] X. Wang *et al.*, "A Weakly-Supervised Framework for COVID-19 Classification and Lesion Localization From Chest CT," *IEEE Trans Med Imaging,* vol. 39, no. 8, pp. 2615-2625, Aug 2020, doi: 10.1109/TMI.2020.2995965.

[67] E.-S. M. El-Kenawy, A. Ibrahim, S. Mirjalili, M. M. Eid, and S. E. Hussein, "Novel feature selection and voting classifier algorithms for COVID-19 classification in CT images," *IEEE Access,* vol. 8, pp. 179317-179335, 2020.

[68] B. Jena, S. Saxena, G. K. Nayak, L. Saba, N. Sharma, and J. S. Suri, "Artificial intelligence-based hybrid deep learning models for image classification: The first narrative review," *Comput Biol Med,* vol. 137, p. 104803, Oct 2021, doi: 10.1016/j.compbiomed.2021.104803.

[69] Y.-H. Wu *et al.*, "Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation," vol. 30, pp. 3113-3126, 2021.

[70] P. P. Phyo and Y.-C. Byun, "Hybrid Ensemble Deep Learning-Based Approach for Time Series Energy Prediction," *Symmetry,* vol. 13, no. 10, p. 1942, 2021. [Online]. Available: https://www.mdpi.com/2073-8994/13/10/1942.

[71] N. A. Alzahab *et al.*, "Hybrid Deep Learning (hDL)-Based Brain-Computer Interface (BCI) Systems: A Systematic Review," *Brain Sci,* vol. 11, no. 1, p. 75, Jan 8 2021, doi: 10.3390/brainsci11010075.

[72] O. S. Albahri *et al.*, "Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects," *J Infect Public Health,* vol. 13, no. 10, pp. 1381-1396, Oct 2020, doi: 10.1016/j.jiph.2020.06.028.

[73] C. Bao, X. Liu, H. Zhang, Y. Li, and J. Liu, "Coronavirus Disease 2019 (COVID-19) CT Findings: A Systematic Review and Meta-analysis," *J Am Coll Radiol,* vol. 17, no. 6, pp. 701-709, Jun 2020, doi: 10.1016/j.jacr.2020.03.006.

[74] Y. S. Kao and K. T. Lin, "A Meta-Analysis of Computerized Tomography-Based Radiomics for the Diagnosis of COVID-19 and Viral Pneumonia," *Diagnostics (Basel),* vol. 11, no. 6, p. 991, May 29 2021, doi: 10.3390/diagnostics11060991.

[75] M. Roberts *et al.*, "Machine learning for COVID-19 detection and prognostication using chest radiographs and CT scans: a systematic methodological review," *arXiv preprint arXiv:.06388,* 2020.

[76] F. Molinari *et al.*, "Hypothesis validation of far-wall brightness in carotid-artery ultrasound for feature-based IMT measurement using a combination of level-set segmentation and registration," *IEEE Transactions on Instrumentation and measurement,* vol. 61, no. 4, pp. 1054-1063, 2012.

[77] Ankush Jamthikar, Deep Gupta, Laura E. Mantella, Luca Saba, Amer M. Johri, and J. S. Suri, "Ensemble Machine Learning and its Validation for Prediction of Coronary Artery Disease and Acute Coronary Syndrome using Focused Carotid Ultrasound," *IEEE Transactions on Instrumentation and Measurement,* 2021, doi: 10.1109/TIM.2021.3139693.

[78] M. Z. Islam, M. M. Islam, and A. Asraf, "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images," *Informatics in medicine unlocked,* vol. 20, p. 100412, 2020.

[79] S. K. Banchhor *et al.*, "Five multiresolution-based calcium volume measurement techniques from coronary IVUS videos: A comparative approach," *computer methods and programs in biomedicine,* vol. 134, pp. 237-258, 2016.

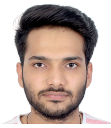
[80] J. S. Suri *et al.*, "Understanding the bias in machine learning systems for cardiovascular disease risk assessment: The first of its kind review," *Computers in Biology and Medicine,* p. 105204, 2022.

[81] S. Das, G. Nayak, L. Saba, M. Kalra, J. S. Suri, and S. Saxena, "An artificial intelligence framework and its bias for brain tumor segmentation: A narrative review," *Computers in Biology and Medicine,* p. 105273, 2022.

[82] S. Paul *et al.*, "Bias Investigation in Artificial Intelligence Systems for Early Detection of Parkinson's Disease: A Narrative Review," *Diagnostics,* vol. 12, no. 1, p. 166, 2022.

Biographies

Jasjit S. Suri is currently Chairman of AtheroPoint, Roseville, CA, USA, dedicated to imaging technologies for cardiovascular and stroke. He received the Director General's Gold medal in 1980. He is a Fellow of AIMBE, APVS, AIUM, SVM, IEEE, and recipient of the Life Time Achievement Award from Marquis (2018). He has co-authored 50 books, 50 patents, has nearly ~22,000 citations with an H-index ~73.



Sushant Agarwal is a member of IEEE and IEEE Computer Society. Currently, he is with Advanced Knowledge Engineering Centre, Global Biomedical Technologies, Inc., Roseville, CA, USA, and AtheroPoint™, USA

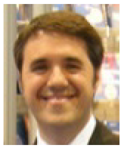
Biswajit Jena received the M.Tech. Degree in Computer Science and engineering with Information Security as specialization from NIT, Rourkela, India, and is currently working toward a Ph.D. degree from the IIIT, Bhubaneswar, India. His research interests are in the field of Medical Image Processing, Machine Learning, and Deep Learning.

Sanjay Saxena, Ph.D., is an Assistant Professor in the Department of Computer Science and Engineering at IIIT, Bhubaneswar, India. He completed his postdoctoral research at Artificial Intelligence in Biomedical Imaging Lab, Perelman School of Medicine, UoP, USA and PhD from IIT BHU, Varanasi, India. His broad area of research is in the implementation of AI techniques in R-n-R studies of cancer".

Ayman El-Baz is a professor, university scholar, and the Chair of the Bioengineering Department at the University of Louisville, Kentucky. El-Baz has 17 years of hands-on experience in the fields of bio-imaging modeling and non-invasive computer-assisted diagnosis systems.

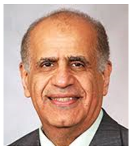
Vikas Agarwal currently works at the Department of Clinical Immunology, Sanjay Gandhi Post Graduate Institute of Medical Sciences. Vikas does research in Clinical Immunology, Clinical Trials and Rheumatology. He has more than 50 international publications to his credit including in NEJM & Oxford Rheumatology.

Mannudeep K. Kalra is an Associate Professor of Radiology, Harvard Medical School Radiologist, Divisions of Thoracic & Cadiovascular Imaging Massachusetts General Hospital. In the past 15 years, he has published more than 170 original research and review articles (h-index of 51 with 10,567 citations, i10 of 144) and 30 book chapters, with most publications being on CT radiation dose optimization.



Luca Saba is with A.O.U. Cagliari, Italy. His research interests are in Multi-Detector-Row Computed Tomography, Magnetic Resonance, Ultrasound, Neuroradiology, and Diagnostic in Vascular Sciences. He has authored ~ 370 peer-review papers and is a frequently invited as guest speaker at RSNA

Klaudija Viskovic is a senior consultant in radiology and a Head of Department of radiology and Ultrasound at the University Hospital for Infectious Diseases in Zagreb, Croatia. Her field of interest is radiological and ultrasound diagnostics of infectious diseases, especially HIV/AIDS. She participated in 8 international projects and presented her research at ~34 national and international conferences.



Mostafa Fatemi is an electrical engineer at the Mayo Clinic in Rochester, Minnesota. Dr. Fatemi's current research areas include ultrasonic methods for tissue viscoelasticity estimation for applications in cancer imaging and bladder function evaluation. He has published extensively in the field of medical ultrasound and holds 10 patents in this field. Dr. Fatemi is an elected Fellow of these institutions: IEEE, AIMBE, ASA, and AIUM.

Subbaram Naidu received Ph.D. degree (1979) in Electrical Engineering (Control Systems Engineering), from Indian Institute of Technology (IIT), Kharagpur, India. He has been nominated for the election to become a Member of the NAE of the United States National Academies Professor Naidu has over ~220 publications including.