IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 November 2022, accepted 3 January 2023, date of publication 9 January 2023, date of current version 12 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3235014

RESEARCH ARTICLE

From Multiple Independent Metrics to Single Performance Measure Based on Objective Function

ASOKE K. NANDI, (Life Fellow, IEEE)
Department of Electronic and Electrical Engineering, Brunel University London, UB8 3PH Uxbridge, U.K.
School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China

e-mail: asoke.nandi@brunel.ac.uk

This work was supported in part by the NSFC, China, through "111 Project" under Grant B20038.

ABSTRACT It is extremely common in engineering to design algorithms to perform various tasks. In data-driven decision making in any field one needs to ascertain the quality of an algorithm. Therefore, a robust assessment of algorithms is essential in deciding the best algorithm as well as in improving algorithms. To perform such an assessment objectively is obvious in the case of a single performance metric, but it is unclear in the case of multiple metrics. Nonetheless, F_1 measure is widely used in cases with two metrics; F_1 measure represents the harmonic mean (*HM*) of two metrics. Of course, there are other means, e.g., the arithmetic mean (*AM*) and the geometric mean (*GM*). As motivations for using them are intuitive and none of them are based on any objective function, it is difficult to judge objectively which is the best one. In this paper, the single metric case is examined to develop two objective functions that are applicable for any number of metrics. These two objective functions lead to two different performance measures – the distance from the origin (*DO*) and the distance from the ideal position (*DIP*). It introduces a new concept of the remaining phase space for the evaluation of the quality of a performance measure. On further and closer examinations of the original goal and the phase space of the metrics, amongst these five measures, either *HM* or *DIP* is found to be the best. Specifically, it is found that *HM* is the best measure at the lower performance end, while *DIP* is clearly the best measure at the higher performance end and is of much practical interest. Rules for deciding the best algorithm and the order of a set of algorithms are presented. These results are derived in the context of multiple independent and bounded metrics. Furthermore, several properties and detailed discussions are provided, following which some published results are reviewed in the present context to elucidate some points.

INDEX TERMS Algorithms, robust assessment, data, decision making, distance from the origin, distance from the ideal position.

I. INTRODUCTION

In data-driven decision making in any field one needs to establish the quality of an algorithm. It is extremely common in engineering and science to design algorithms to perform various tasks [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. For example, either in detection or classification using machine learning, one develops and compares many different algorithms. Therefore, it is essential to design a robust performance metric to measure the performance of algorithms. Without this, it is not possible either to choose the best/better algorithm or to gauge the improvement of an algorithm.

In the case of a single performance metric, it is clear that the best algorithm is the one corresponding to the largest value of the performance metric, when 0 represents the worst performance and 1 represents the best performance. That said, it is unclear how to make this judgement in the case of multiple metrics. Nonetheless, such decisions are made; for example, F_1 measure is widely used in cases with two metrics, which is essentially the harmonic mean (*HM*) of these two metrics. Of course, there are other means, e.g.,

The associate editor coordinating the review of this manuscript and approving it for publication was Paul D. Yoo.

the arithmetic mean (*AM*) and the geometric mean (*GM*). As motivations for using them are intuitive and not coming from any objective function, it is difficult to judge objectively which is the best one of these three.

In this paper it is assumed that every performance metric is independent, important, and bounded. Without any loss of generalisation, each metric is considered to be bounded between 0 and 1, where 0 represents the worst performance and 1 represents the best performance. The single metric case is studied to develop two objective functions that offer identical results in the single metric case and can easily be extended to any number of metrics. These two objective functions give rise to two different performance measures – the distance from the origin (*DO*) and the distance from the ideal position (*DIP*).

In the multiple-metric cases, it is always true that $DO \geq AM \geq GM \geq HM$, the equality holds only when all metric values are equal. Furthermore, in the phase space of multiple-metric cases, it is proven in this paper that $DO \geq AM \geq DIP$ is always true, with the equality being valid only when all metric values are equal. On further and closer examinations of the original goal and the phase space of the metrics, amongst these five measures it is demonstrated that *HM* is the best measure at the lower performance end, but at the higher performance end, which is of much practical interest, *DIP* is clearly the best measure.

Three major contributions of this paper are considered to be

1) It addresses a much-required knowledge gap in objective evaluation of performance measures. It is considered to be timely and is of relevance to many different fields, including ones not illustrated in the paper.
2) It introduces a new concept of the remaining phase space for the evaluation of the quality of a performance measure.
3) It proposes two new performance measures, namely *DO* and *DIP*. The *DIP* turns out to be the best amongst the five measures examined in this paper.

In Section II, three existing performance measures are discussed. Two objective functions are developed leading to two new performance measures which are proposed in Section III. Several properties of these five measures, exemplifying their similarities and differences, are presented in Section IV. Relative ranking of algorithms, which is the ultimate goal, is discussed in section V. Several published results are reviewed and compared using these five measures in Section VI to elucidate some points about these measures. Discussions in Section VII offer more insights into these measures as well as how they may be extended if all metrics are not equally important. Finally, conclusions are presented in Section VIII.

II. PERFORMANCE MEASURES

To measure the performance of an algorithm it is essential to have a performance metric. In this paper it is assumed that

1) every performance metric is bounded,
2) every performance metric is independent, and
3) every performance metric is important.

Whilst every metric is bounded, individually they do not have to have the same bounds, since each metric is normalised such that each is bounded between 0 and 1, where 0 represents the worst performance and 1 represents the best performance. It is further assumed that these metrics are equally important, although this assumption can be relaxed as outlined in Section VII. When one is using only one metric it is easy to compare the performance of several algorithms. The algorithm with the largest value of the metric is judged to be the best amongst them. When there are several metrics, it has not been so straightforward. For example, consider that there are three algorithms (namely, A_1, A_2, and A_3) and there is only one metric (namely, m_1). Denote the value of the metric m_j of the algorithm A_i as $m_j(A_i)$. Let $m_1(A_i)$ be 0.59, 0.59, and 0.60 for i = 1, 2, and 3. In this case, one would judge the algorithm A_3 to be the best.

Now, consider that there are the same three algorithms (namely, A_1, A_2, and A_3), but use a different metric (namely, m_2). Let $m_2(A_i)$ be 0.93, 0.91, and 0.90 for i = 1, 2, and 3. In this case, one would judge the algorithm A_1 to be the best. So, A_3 is found to be the best using only m_1, while A_1 is found to be the best using only m_2. Which is the best algorithm if one uses both metrics? More on this can be found in subsection VI-A. In the following are three subsections reviewing three performance measures.

A. ARITHMETIC MEAN

The arithmetic mean (*AM*) is defined as

$$AM = \frac{1}{N}\sum_{i=1}^{N} m_i \tag{1}$$

When there are several independent metrics, the *AM* provides a score between the largest and the smallest metric values. Also note that the value of *AM* is bounded between 0 (representing the worst possible value) and 1 (representing the best possible value). Thus, the algorithm with the largest value of *AM* can be considered the best. Although the use of *AM* is not so common, a recent example of its use can be found in [14]. The motivation for the use of *AM* appears to be related to statistical averaging, but beyond that its relevance has not been discussed in the literature.

B. HARMONIC MEAN

The harmonic mean (*HM*) is defined as

$$HM = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i}\right)^{-1} \tag{2}$$

When there are several independent metrics, the *HM* provides a score between the largest and the smallest metric values. Also note that the value of *HM* is bounded between 0 (representing the worst possible value) and 1 (representing the best possible value). Thus, the algorithm with the largest value

of *HM* can be considered the best. The use of *HM* is very common. Some recent examples of its use can be found in [6], [7], [8], [9], and [10].

In the two metrics scenario, *HM* is popularly referred to as the F_1 score. The F-measure was introduced by Chinchor in the context of measuring the performance of message understanding systems [15]. The two metrics for this F-measure were precision and recall.

For motivation, Chinchor wrote, ''The F-measure is higher if the values of recall and precision are more towards the center of the precision-recall graph than at the extremes and their sums are the same. So ...a system which has recall of 50% and precision of 50% has a higher F-measure than a system which has recall of 20% and precision of 80%. This behaviour is exactly what we want from a single measure.'' [15].

In [16] Sasaki wrote, ''The harmonic mean is more intuitive than the arithmetic mean when computing a mean of ratios. Suppose that you have a finger print recognition system and its precision and recall be 1.0 and 0.2, respectively. Intuitively, the total performance of the system should be very low because the system covers only 20% of the registered finger prints, which means it is almost useless. The arithmetic mean of 1.0 and 0.2 is 0.6 whereas the harmonic mean of them is 1/3. As you see in this example, the harmonic mean (0.333...) is a more reasonable score than the arithmetic mean.''

The above comments can be abstracted as follows:

Notion 1: When comparing different measures, a smaller value of a measure indicates a better measure.

The precursor to F-measure is the E-measure of van Rijsbergen [17], [34]. Essentially, $E = 1 - F$. Van Rijsbergen also wrote about ''an intuitive way of measuring'' before introducing the E-measure. Thus, the motivation for the use of *HM* is more of an intuitive expectation and beyond that its relevance has not been elucidated. Indeed, the same author wrote, ''The preceding argument in itself is not sufficient to justify the use of this particular composite measure''.

C. GEOMETRIC MEAN

The geometric mean (*GM*) is defined as

$$GM = \left(\prod_{i=1}^{N} m_i\right)^{\frac{1}{N}} \tag{3}$$

When there are several independent metrics, the *GM* provides a score between the largest and the smallest metric values. Also note that the value of *GM* is bounded between 0 (representing the worst possible value) and 1 (representing the best possible value). Thus, the algorithm with the largest value of *GM* can be considered the best. We have not noticed any instances of the use of *GM* in such contexts. The motivation for the use of *GM* can be built around the fact that it provides a score between the lowest and the highest values, but its appropriateness has not been clarified.

It should be noted that GM has similar properties to *HM* or F_1 measure. For example, like the desirable property of F-measure as commented by Chinchor [15], *GM* values are also higher if the values of recall and precision are more towards the centre of the precision-recall graph than at the extremes and their sums are the same. Also, as considered reasonable by Sasaki [16] of the *HM* measure, *GM* values are lower than *AM* values.

III. PROPOSED MEASURES

In section II the three measures, namely *AM*, *GM*, and *HM*, have been outlined. Of these, *HM* is widely used and the most popular. There are intuitive motivations for each of them, but these are not based on any cost functions.

In the rest of this section, the one metric scenario is explored to develop two cost functions, leading to two new measures. One of these new measures appears to be very appropriate in this context.

A. DISTANCE FROM THE ORIGIN

Remember the assumptions in this paper are that every performance metric is bounded and they are independent. Moreover, each metric is normalised such that each is bounded between 0 and 1, where 0 represents the worst performance and 1 represents the best performance.

A single metric represents a segment of a straight line between 0 and 1, and the largest value of this metric can be construed as the maximum distance from the origin (*DO*), i.e., from 0 to its value. Thus, one can maximise the distance from the origin. For the case of a single metric, both the largest value and the maximum distance from the origin give the same result, i.e., they are essentially the same.

In multiple metric cases, the distance from the origin can be expressed as $\sqrt{\sum_{i=1}^{N} m_i^2}$. When there are more than one (e.g., N) metrics, then the space of metrics has N dimensions. While the maximum possible distance in one dimension is 1, it is $\sqrt{N}$ in this N-dimensional space. All the aforementioned measures – *AM*, *GM*, and *HM* – are bounded between 0 and 1. To ensure the same property for *DO*, it is proposed that

$$DO = \frac{\sqrt{\sum_{i=1}^{N} m_i^2}}{\sqrt{N}} \tag{4}$$

When there are several independent metrics, the *DO* provides a score between the largest and the smallest metric values. Also note that the value of *DO* is bounded between 0 (representing the worst possible value) and 1 (representing the best possible value). Thus, the algorithm with the largest value of *DO* can be considered the best. There have been no references in the literature of *DO* in the context of multiple metrics. The motivation for the use of *DO* is built around the fact that it provides a score between the lowest and the highest values, and, more importantly, it is based on a cost function that is appropriate and in common use in a single metric scenario.

B. DISTANCE FROM THE IDEAL POSITION

Remember that each metric is normalised such that it is bounded between 0 and 1, where 0 represents the worst performance and 1 represents the best performance.

As a single metric represents a segment of a straight line between 0 and 1, the largest value of this metric can be construed as the maximum distance from the origin (*DO*), i.e., from 0 to its value. Since the ideal value is 1, one can alternatively formulate the best being the minimum distance from the ideal value of 1. Thus, one can minimise the distance from the ideal position. For the case of a single metric, the largest value and the maximum distance from the origin as well as the minimum distance from the ideal position (*DIP*) give the same result, i.e., all three produce the same outcome.

In multiple metric cases, the distance from the ideal position can be written as $\sqrt{\sum_{i=1}^{N}(1-m_i)^2}$. When there are more than one (e.g., N) metrics, then the space of metrics has N dimensions. While the maximum possible distance in one dimension is 1, it is $\sqrt{N}$ in this N-dimensional space. All the aforementioned measures – *AM*, *GM*, *HM*, and *DO* – are bounded between 0 (worst) and 1 (ideal). To keep the same property for *DIP*, it is proposed that

$$DIP = 1 - \frac{\sqrt{\sum_{i=1}^{N}(1-m_i)^2}}{\sqrt{N}} \tag{5}$$

When there are several independent metrics, the *DIP* provides a score between the smallest and the largest metric values. Also note that the value of *DIP* is bounded between 0 (representing the worst possible value) and 1 (representing the best possible value). Thus, the algorithm with the largest value of *DIP* can be considered the best. Although there have been no formulations in the literature of *DIP* in such contexts, the motivation for the use of *DIP* is built around the fact that it provides a score between the lowest and the highest values, and, more importantly, it is based on a cost function that is more appropriate than the rest, in that it aims to measure the nearness to the ideal position which is the ultimate goal.

IV. PROPERTIES

There are seven subsections below exploring several properties of the aforementioned five measures, including the two proposed measures - *DO* and *DIP*. These compare and contrast the five measures. There are some graphical representations included in the following. Without any loss of generality, much of the discussions in this section will be in the contexts of two metrics as their graphical representations will be better appreciated on two dimensional plots. When considering only two metrics, m_1 will be labelled as x while m_2 will be labelled as y.

A. ALL METRIC VALUES ARE EQUAL

Consider the scenario that $m_1 = m_2 = \ldots = m_N = m$. Then

$$AM = \frac{1}{N}\sum_{i=1}^{N} m_i = m$$

$$GM = \left(\prod_{i=1}^{N} m_i\right)^{\frac{1}{N}} = m$$

$$HM = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i}\right)^{-1} = m$$

$$DO = \frac{\sqrt{\sum_{i=1}^{N} m_i^2}}{\sqrt{N}} = m$$

$$DIP = 1 - \frac{\sqrt{\sum_{i=1}^{N}(1-m_i)^2}}{\sqrt{N}} = m$$

Therefore, when all metric values are equal, all the five measures give the same value, i.e., at every point on the diagonal on the $x-y$ plane from (0,0) to (1,1), all measure values are equal.

B. SUM OF TWO METRIC VALUES ARE CONSTANT

In this scenario, let $m_1 = (m-z)$ and $m_2 = (m+z)$, such that m is fixed but z is not zero and variable, and $(m_1+m_2) = 2m$. Then

$$AM(m-z, m+z) = \frac{1}{2}\sum_{i=1}^{2} m_i = m$$

Thus, $AM\ (m, m) = AM(m-z, m+z)$.

Now,

$$GM(m-z, m+z) = \left(\prod_{i=1}^{2} m_i\right)^{\frac{1}{2}} = \sqrt{((m-z)\,(m+z))}$$
$$= \sqrt{m^2 - z^2} < m$$

Therefore, $GM\ (m, m) > GM(m-z, m+z)$.

Also,

$$HM(m-z, m+z) = \left(\frac{1}{2}\sum_{i=1}^{2}\frac{1}{m_i}\right)^{-1} = \frac{2\,(m-z)\,(m+z)}{2m}$$
$$= \frac{m^2 - z^2}{m} < m$$

Thus, $HM\ (m, m) > HM(m-z, m+z)$.

Now,

$$DO(m-z, m+z) = \frac{\sqrt{\sum_{i=1}^{2} m_i^2}}{\sqrt{2}} = \frac{\sqrt{(m-z)^2 + (m+z)^2}}{\sqrt{2}}$$
$$= \sqrt{m^2 + z^2} > m$$

Therefore, $DO\ (m, m) < DO(m-z, m+z)$.

Finally,

$$DIP(m-z, m+z) = 1 - \frac{\sqrt{\sum_{i=1}^{2}(1-m_i)^2}}{\sqrt{2}}$$
$$= 1 - \frac{\sqrt{(1-m+z)\,(1-m-z)}}{\sqrt{2}}$$
$$= 1 - \sqrt{(1-m)^2 + z^2}$$

Since $((1-m)^2+z^2) > (1-m)^2$,
$DIP(m-z, m+z) < (1-(1-m)) = m$. Hence,

$$DIP\,(m, m) > DIP(m-z, m+z).$$

It may be helpful to consider some geometrical aspects. On the $m_1 - m_2$ plane, the point described by $m_1 = m_2 = m$ lie on the main diagonal between (0,0) and (1,1). On the other hand, the line described by the equation $m_1 + m_2 = 2m$ is perpendicular to the main diagonal with the two lines intersecting at the point (m, m).

At the intersecting point (m, m), all five measures give the same value. As one moves away from this intersecting point along the perpendicular line $(m_1 + m_2 = 2m)$, *DO* values increase monotonically and AM values remain constant, while *GM* values, *HM* values, and *DIP* values decrease monotonically. In this respect, *GM*, *HM*, and *DIP* behave similarly.

Remember that Chinchor wrote, "The F-measure is higher if the values of recall and precision are more towards the center of the precision-recall graph than at the extremes and their sums are the same. . . . This behaviour is exactly what we want from a single measure." [15]. What is now observed is that *GM* and *DIP* have the same behaviour as the F-measure in this respect. Therefore, in the spirit of the above reasoning, *GM* and *DIP* should be considered further as well.

C. TWO-DIMENSIONAL PHASE SPACE

In this subsection, comparisons are made of the five measures in the contexts of only two metrics, through their graphical representations. This restriction to two metrics is helpful for portrayals on two dimensional plots. Now the metric m_1 is labelled as x while the metric m_2 is labelled as y.

To compare the five measures in the contexts of two metrics, the values of each of these measures are set to some fixed value, f. For each measure, the curve corresponding to the same fixed value is calculated.

For *AM*, $x + y = 2f$. This represents a straight line on the x-y plane, with a slope of -1 and an intercept of $2f$. This is presented as the black line in Figure 1 for $f = 0.72$.

For *GM*, $\sqrt{xy} = f$. Thus, $y = f^2/x$. This is not a straight line. The corresponding curve on the x-y plane is shown as the magenta curve in Figure 1 for $f = 0.72$.

For *HM*, $\left(\frac{1}{2}(\frac{1}{x}+\frac{1}{y})\right)^{-1} = f$. Or, $x+y = 2xy/f$. Thus, $y = x\left(\frac{2x}{f}-1\right)^{-1}$. This is not a straight line. The corresponding curve on the x-y plane is depicted as the blue curve in Figure 1 for $f = 0.72$.

For *DO*, $\sqrt{x^2+y^2}/\sqrt{2} = f$. This represents a circle around the centre at (0,0) of radius $\sqrt{2}\,f$, and it can be rewritten as $y = \sqrt{2f^2 - x^2}$. The corresponding curve on the x-y plane is drawn as the red curve in Figure 1 for $f = 0.72$.

For *DIP*, $1 - \frac{\sqrt{(1-x)^2+(1-y)^2}}{\sqrt{2}} = f$. This can be rewritten as $(1-x)^2 + (1-y)^2 = 2(1-f)^2$, which represents a circle around the centre at (1,1) of radius $\sqrt{2}(1-f)$. The corresponding curve on the x-y plane is displayed as the green curve in Figure 1 for $f = 0.72$.

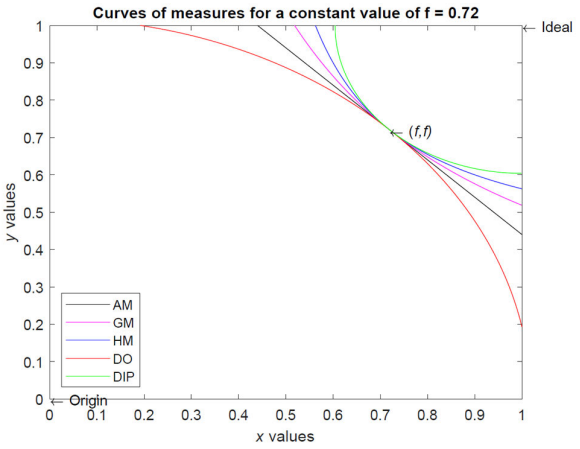

FIGURE 1. For each of the five measures the curve of a measure is depicted for a constant value of $f = 0.72$.

Figure 1 depicts the curves for each of the five measures for the same fixed value of 0.72. The horizontal axis is the x-axis and the vertical axis is the y-axis. The bottom left corner is the origin and the top right corner represents the ideal values of the metrics. The curves represent *DO* in red, *AM* in black, *GM* in magenta, *HM* in blue, and *DIP* in green. All these curves are different, though they all intersect at $(x, y) = (f, f)$. This is a pictorial verification of the theoretical results in subsection IV-A. Also, the theoretical deductions in subsection IV-B concerning what happens as one moves away from this intersecting point along the perpendicular line $(x + y = 2f)$ are depicted in Figure 1.

Remember that the ideal position of $(x, y) = (1, 1)$. It is important to note that, for a fixed measure value of 0.72, *DO* allows a more varied range combinations of (x, y) that are further away from the ideal position than *AM*, which allows a more varied range combinations of (x, y) that are further away from the ideal position than *GM*, which allows a more varied range combinations of (x, y) that are further away from the ideal position than *HM*, which allows a more varied range combinations of (x, y) that are further away from the ideal position than *DIP*.

Figure 1 presented five curves corresponding to the constant f value of 0.72 (i.e., in the higher performance region), for each of the five measures. In contrast, curves corresponding to a different constant f value of 0.35, in the lower performance region, for each of the five measures are displayed in Figure 2 to evince different relationships between *HM* and *DIP* curves. The following facts are noted:

1) The relative positions of *DO*, *AM*, *GM*, and *HM* curves in Figure 1 and Figure 2 are the same, i.e., the remaining phase space of *HM* < the remaining phase space of *GM* < the remaining phase space of *AM* < the remaining phase space of *DO*. It can be shown that this is not only true for these two f values, but also it is true for all f values.

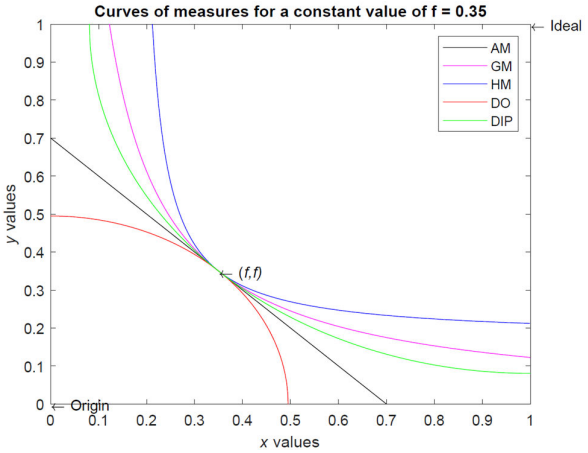

FIGURE 2. For each of the five measures the curve of a measure is presented for a constant value of $f = 0.35$.

2) The relative positions of *DIP* curve and *HM* curve in Figure 1 and Figure 2 are different. In Figure 1 the remaining phase space of *DIP* < the remaining phase space of *HM*, while in Figure 2 the remaining phase space of *HM* < the remaining phase space of *DIP*.
3) Exploring the above observations further, it can be demonstrated that the remaining phase space of *HM* is equal to the remaining phase space of *DIP* around $f = 0.63$ (see section IV-F). Thus, the remaining phase space of *HM* < the remaining phase space of *DIP* for $f < 0.63$, while the remaining phase space of *DIP* < the remaining phase space of *HM* for $f > 0.63$ (see section IV-F).

D. CURVATURES

It is observed from Figures 1 and 2 that the curves of the five measures corresponding to the same fixed value of the measures have different shapes at the intersection point $(x, y) = (f, f)$. Below the curvatures of these five curves are explored. The curvature, K, can be written as

$$K = \frac{\left|\frac{d^2y}{dx^2}\right|}{\left[1 + \left(\frac{dy}{dx}\right)^2\right]^{\frac{3}{2}}} \tag{6}$$

For *AM*, $x + y = 2f$. This leads to $\frac{dy}{dx} = -1$ and $d^2y/dx^2 = 0$. As this is a straight line, it has no curvature ($K = 0$).

For *GM*, $\sqrt{xy} = f$. Or, $\frac{1}{2}xdy + \frac{1}{2}ydx = 0$. Or, $dy/dx = -y/x$. Thus, $\frac{d^2y}{dx^2} = \frac{2y}{x^2} = \frac{2}{f} > 0$ and $dy/dx = -1$ at the intersection point $(x, y) = (f, f)$. Hence, the curvature is $1/(\sqrt{2}f)$. Since, $\frac{d^2y}{dx^2} > 0$, it is concave upward.

For *HM*, $x + y = 2xy/f$. Or, $dx + dy = (2ydx + 2xdy)/f$. Or, $dy/dx = (2y - f)/(f - 2x)$. Thus, $\frac{d^2y}{dx^2} = \left(\frac{2y-f}{(f-2x)^2}\right)(2) + \frac{1}{(f-2x)}(2\frac{dy}{dx}) = \frac{4(2y-f)}{(f-2x)^2} = \frac{4}{f} > 0$ and $dy/dx = -1$ at the intersection point $(x, y) = (f, f)$. Hence, the curvature is $2/(\sqrt{2}f)$. Since, $\frac{d^2y}{dx^2} > 0$, it is concave upward.

TABLE 1. Some observations on each of the five curves corresponding to the constant measure value, f, at the intersection point $(x,y) = (f,f)$.

Measure	*DO*	*AM*	*GM*	*HM*	*DIP*
d^2y/dx^2 at (f,f)	$-2/f$	0	$2/f$	$4/f$	$2/(1-f)$
Curvature at (f,f)	$\frac{1}{\sqrt{2}f}$	0	$\frac{1}{\sqrt{2}f}$	$\frac{2}{\sqrt{2}f}$	$\frac{1}{\sqrt{2}(1-f)}$
Shape of curve of constant measure value f at (f,f)	Concave downward	Flat	Concave upward	Concave upward	Concave upward

For *DO*, $\sqrt{x^2+y^2}/\sqrt{2} = f$. So, $x^2 + y^2 = 2f^2$ and $2xdx + 2ydy = 0$. Or, $dy/dx = -x/y$. Thus, $\frac{d^2y}{dx^2} = -\frac{1}{y} + \frac{x}{y^2}\frac{dy}{dx} = -\frac{1}{y} - \frac{x^2}{y^3} = -\frac{x^2+y^2}{y^3} = -\frac{2f^2}{y^3} = -\frac{2}{f} < 0$ and $dy/dx = 1$ at the intersection point $(x, y) = (f, f)$. Hence, the curvature is $1/\sqrt{2}f$. Since, $\frac{d^2y}{dx^2} < 0$, it is concave downward.

For *DIP*, $(1-x)^2 + (1-y)^2 = 2(1-f)^2$. Therefore, $2(1-x)dx + 2(1-y)dy = 0$. Or, $dy/dx = -(1-x)/(1-y)$. Thus, $\frac{d^2y}{dx^2} = \frac{1}{1-y} - \frac{1-x}{(1-y)^2}\frac{dy}{dx} = \frac{1}{1-y} + \frac{(1-x)^2}{(1-y)^3} = \frac{(1-x)^2+(1-y)^2}{(1-y)^3} = \frac{2(1-f)^2}{(1-y)^3} = \frac{2}{1-f} > 0$ and $dy/dx = -1$ at the intersection point $(x, y) = (f, f)$. Therefore, the curvature is $1/(\sqrt{2}(1-f))$. Since, $\frac{d^2y}{dx^2} > 0$, it is concave upward.

Some of these observations on each of the five curves corresponding to the constant measure value, f, at the intersection point $(x, y) = (f, f)$ are presented in Table 1.

Few comments follow:

1) It is clear that, at the intersection point $(x, y) = (f, f)$, the curvature of *DO* is concave downward while the same for each of *GM*, *HM*, and *DIP* is concave upward. Considering that the ideal position is at (1, 1), measures whose curves are concave upward are desirable.
2) At the intersection point $(x, y) = (f, f)$, the curvature of *GM* is $1/\sqrt{2}f$ and the curvature of *HM* is $2/\sqrt{2}f$. Independent of the value of f, the curvature of *HM* is twice as large as that of *GM*, implying that the *HM* curve is closer to the ideal position. More on this can be found in sections IV-F and IV-G.
3) At the intersection point $(x, y) = (f, f)$, the curvature of *DIP* is $1/(\sqrt{2}(1-f))$. The relative values of curvatures of *DIP*, *HM*, and *GM* depend on the value of f. For higher values of f, the curvature of *DIP* is greater than that of *GM* and *HM*. For lower values of f, the curvature of *DIP* is smaller than that of *HM* but greater than *GM*, while, for even lower values of f, the curvature of *DIP* is smaller than that of *HM* and *GM*. More along this line be found in sections IV-F and IV-G.

E. TANGENT AM

Theorem 1: AM is a tangent to *GM* at the intersection point $(x, y) = (f, f)$.

Proof: For *GM*,

$$\sqrt{xy} = f$$
$$\frac{1}{2}xdy + \frac{1}{2}ydx = 0$$
$$\frac{dy}{dx} = -\frac{y}{x}$$

At the intersection point $(x, y) = (f, f)$, $dy/dx = -1$. Or, $y = -x+c$, where $c = 2f$ as it passes through the intersection point. Thus, the equation of the tangent to *GM* is $y = -x+2f$, which is the *AM* line represented by the black line in Figure 1. **QED**.

Theorem 2: AM is a tangent to *HM* at the intersection point $(x, y) = (f, f)$.

Proof: For *HM*,

$$x + y = \frac{2xy}{f}$$
$$dx + dy = \frac{2ydx + 2xdy}{f}$$
$$\frac{dy}{dx} = \frac{2y - f}{f - 2x}$$

At the intersection point $(x, y) = (f, f)$, $dy/dx = -1$. Or, $y = -x+c$, where $c = 2f$ as it passes through the intersection point. Thus, the equation of the tangent to *HM* is $y = -x+2f$, which is the *AM* line represented by the black line in Figure 1. **QED**.

Theorem 3: AM is a tangent to *DO* at the intersection point $(x, y) = (f, f)$.

Proof: For *DO*,

$$x^2 + y^2 = 2f^2$$
$$2xdx + 2ydy = 0$$
$$\frac{dy}{dx} = -\frac{x}{y}$$

At the intersection point $(x, y) = (f, f)$, $dy/dx = -1$. Or, $y = -x+c$, where $c = 2f$ as it passes through the intersection point. Thus, the equation of the tangent to *DO* is $y = -x+2f$, which is the *AM* line represented by the black line in Figure 1. **QED**.

Theorem 4: AM is a tangent to *DIP* at the intersection point $(x, y) = (f, f)$.

Proof: For *DIP*,

$$(1 - x)^2 + (1 - y)^2 = 2(1 - f)^2$$
$$2(1 - x)dx + 2(1 - y)dy = 0$$
$$\frac{dy}{dx} = -\frac{1 - x}{1 - y}$$

At the intersection point $(x, y) = (f, f)$, $dy/dx = -1$. Or, $y = -x+c$, where $c = 2f$ as it passes through the intersection point. Thus, the equation of the tangent to *DIP* is $y = -x+2f$, which is *AM* represented by the black line in Figure 1. **QED**.

In summary, the line of constant *AM* is the tangent to each of the curves of constant *GM*, *HM*, *DO*, and *DIP* at the intersection point $(x, y) = (f, f)$ along the diagonal line.

F. AREAS ABOVE THE CURVES

The ideal position in Figure 1 is (1,1). So, it is instructive to calculate how much of the phase space (i.e., area on the *x*-*y* plane) remains between each of these five curves and the ideal position when each of these measures are set to some fixed value, f. As one reaches the ideal position, no phase space will be left over. Thus, the least remaining phase space (i.e., area) is desirable. In the following calculations, without any loss of generality, it is considered that $f \geq 1/\sqrt{2}$.

Notion 2: The measure that leaves the least remaining phase space between each of these five curves and the ideal position when it is set to some fixed value, *f*, is considered the best.

For *AM*, $x + y = 2f$. Thus, the remaining phase space between this *AM* straight line and the ideal position is a right-angled triangle with corners at $(2f - 1, 1)$, $(1, 1)$, and $(1, 2f - 1)$. Therefore, the remaining phase space is $2(1 - f)^2$.

For *GM*, $\sqrt{xy} = f$. Or, $y = f^2/x$. Here, the remaining phase space is the area between the *GM* arc and the ideal position defined by the three points at $(f^2, 1)$, $(1, 1)$, and $(1, f^2)$. It can be shown that the remaining phase space is $\left(\left(1 - f^2\right) + 2f^2 \ln(f)\right)$.

For *HM*, $x + y = 2xy/f$. Now, the remaining phase space is the area between the *HM* arc and the ideal position defined by the three points at $(f/(2 - f), 1)$, $(1, 1)$, and $(1, f/(2 - f))$. It can be demonstrated that the remaining phase space is $\left(1 - f - \left(\frac{f^2}{2}\right)(\ln(2 - f) - \ln(f))\right)$.

For *DO*, $\sqrt{x^2 + y^2}/\sqrt{2} = f$. Or, $x^2 + y^2 = 2f^2$. So, the remaining phase space is the area between the *DO* circular arc and the ideal position defined by the three points at $(\sqrt{2f^2 - 1}, 1)$, $(1, 1)$, and $(1, \sqrt{2f^2 - 1})$. It can be shown that the remaining phase space is $\left(1 - \sqrt{2f^2 - 1} - \vartheta f^2\right)$, where $\vartheta = 2\sin^{-1}\left(\sqrt{f^2 - \sqrt{2f^2 - 1}}/\left(\sqrt{2}f\right)\right)$.

For *DIP*, $(1 - x)^2 + (1 - y)^2 = 2(1 - f)^2$. In this case, the remaining phase space is the area between the *DIP* circular arc and the ideal position, which is one quarter of a circle of radius $(\sqrt{2} - \sqrt{2}f)$. Therefore, the remaining phase space is $(\frac{\pi}{4})(\sqrt{2} - \sqrt{2}f)^2 = \pi(1 - f)^2/2$.

Figure 3 depicts the remaining phase space (i.e., area on the *x*-*y* plane) between each of these five curves and the ideal position when each of these measures are set to some fixed value, f. Note that the maximum possible phase space is 1. The horizontal axis represents the f-values while the vertical axis represents the remaining phase spaces. For any fixed value of f, *DO* (red) has the most phase space left, *AM* (black) has less phase space left, *GM* (magenta) has lesser phase space left, *HM* (blue) has even less phase space left, and *DIP* (green) has the least phase space left. Thus, in this range of f values, DIP guarantees the smallest phase space as well as is nearest to the ideal position.

It is clear from Figure 3 that the remaining phase spaces for all five measures go to 0 as f goes to 1. Although this is not surprising since the ideal position is reached when f goes

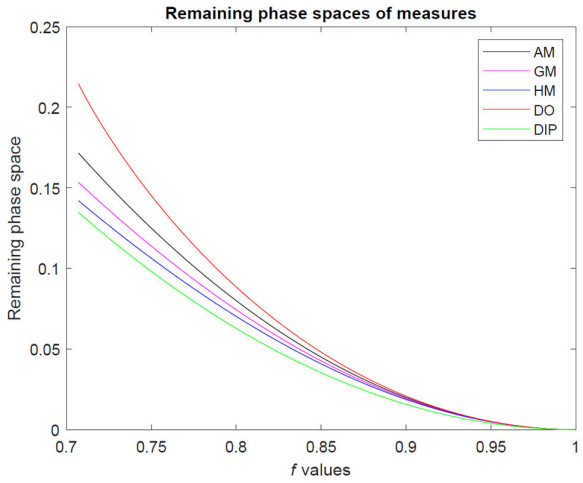

FIGURE 3. For each of the five measures the remaining phase space between the curve of a measure for a constant value of f and the ideal position versus f in the range (0.72, 1) is displayed.

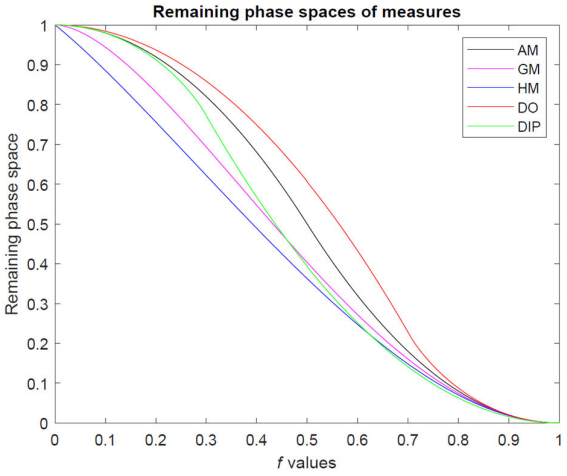

FIGURE 5. For each of the five measures the remaining phase space between the curve of a measure for a constant value of f and the ideal position versus f in the range (0,1) is presented.

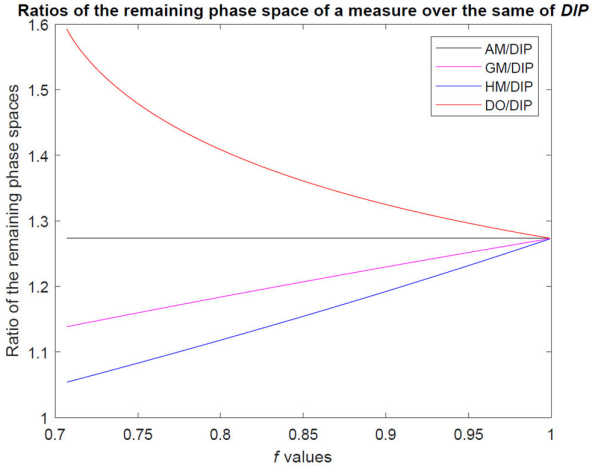

FIGURE 4. For each of the five measures the ratio of the remaining phase space between the curve of a measure for a constant value of f and the ideal position over the same of DIP versus f in the range (0.72, 1) is shown.

to 1, something interesting is observed when one considers the ratios of the remaining phase spaces of four measures with respect to *DIP* as f goes to 1. In Figure 4, the horizontal axis represents the f-values while the vertical axis represents the ratios of remaining phase spaces. It displays the ratio of the remaining phase spaces of *DO* and *DIP* in red, the ratio of the remaining phase spaces of *AM* and *DIP* in black, the ratio of the remaining phase spaces of *GM* and *DIP* in magenta, and the ratio of the remaining phase spaces of *HM* and *DIP* in blue. The first observation is that, as f tends to 1, all four ratios converge to the same value, but it is different from 1.

For *AM*, the remaining phase space is $2\,(1-f)^2$. Let $f = (1-\varepsilon)$. Then f tends to 1, as ε goes to 0. In this case, $2\,(1-f)^2 = 2\varepsilon^2$. For *GM*, the remaining phase space, as f tends to 1, is $\left(\left(1-f^2\right)+2f^2\ln(f)\right) = \left(1-(1-\varepsilon)^2\right) + 2\,(1-\varepsilon)^2\ln(1-\varepsilon) \approx 2\varepsilon^2$, as ε goes to 0. For *HM*, the remaining phase space between the blue curve and the ideal position is $\left(1-f-\left(\frac{f^2}{2}\right)(\ln(2-f)-\ln(f))\right) \approx 2\varepsilon^2$, as f tends to 1, which is equivalent to ε going to 0. For *DO*, the remaining phase space is $\left(1-\sqrt{2f^2-1}-\vartheta f^2\right)$, where $\vartheta = 2\sin^{-1}\left(\sqrt{f^2-\sqrt{2f^2-1}}/\left(\sqrt{2}\,f\right)\right) \approx 2\varepsilon^2$, as ε tends to 0, i.e., f tends to 1. For *DIP*, the remaining phase space is $\pi\,(1-f)^2/2 = \pi\varepsilon^2/2$. It is clear that the values of each of *AM*, *GM*, *HM*, and *DO* tends to the same value of $2\varepsilon^2$, except that of DIP which tends to $\pi\varepsilon^2/2$. Thus, the ratios of the remaining phase space of *AM*, *GM*, *HM*, and *DO* over *DIP* are the same at $2\varepsilon^2/(\pi\varepsilon^2/2)$, which is equal to $4/\pi$ and is different from 1.

The second observation is that the ratio of the remaining phase spaces of *AM* and *DIP* appears to be constant, i.e., independent of the value of f. Indeed, this ratio is $(2\,(1-f)^2)/(\pi\,(1-f)^2/2) = 4/\pi = 1.273$, and this is independent of the value of f as long as $f \geq 1/\sqrt{2}$.

In Figure 3, the remaining phase spaces for all five measures, for f values between 0.72 and 1, have been displayed in the higher performance end. To appreciate the whole phase space, i.e., lower performance ($f < 0.63$) and higher performance ($f > 0.63$) regions, the remaining phase spaces for all five measures for f values in the complete range of 0 to 1 are displayed in Figure 5. It is clear that *HM* is the best in the lower performance region and that *DIP* is the best in the higher performance region.

G. RELATIVE VALUES

For any pair of metrics (x, y), it can be proven that the value of *HM* $\leq$ the value of *GM* $\leq$ value of *AM* $\leq$ the value of *DO*. For example, $(x+y)^2 - 4xy = (x-y)^2 \geq 0$. Therefore, $4AM^2 - 4GM^2 \geq 0$ and $AM \geq GM$. Also, $(x+y)^2 - 2xy = x^2+y^2 \geq 0$. Or, $2GM^4/HM^2 - 2GM^2 \geq 0$. Or, $GM^2/HM^2 - 1 \geq 0$. Thus, taking the positive root, $GM \geq$

HM. Now, $\left(2x^2+2y^2\right)-\left(x^2+y^2+2xy\right)=(x-y)^2\geq 0$. Or, $4DO^2-4AM^2\geq 0$. Taking the positive root, one finds that $DO\geq AM$. This concludes the proof, in two metrics cases, that $DO\geq AM\geq GM\geq HM$ and the equality sign applies when $x=y$.

It well known that the result in above paragraph is true more generally than the 2-metric cases.

Theorem 5: $DO\geq AM$ is true in multiple metric cases.

Proof: Now,

$$AM=\frac{1}{N}\sum_{i=1}^{N}m_i=m$$

Let $m_i=m+z_i$, for $i=1,\ldots,N$. Then $\sum_{i=1}^{N}z_i=0$. Thus,

$$\begin{aligned}N\,(DO)^2&=\sum_{i=1}^{N}m_i^2=\sum_{i=1}^{N}(m+z_i)^2\,,\\&=Nm^2+\sum_{i=1}^{N}z_i^2+2m\sum_{i=1}^{N}z_i,\\N\,(DO)^2&=Nm^2+\sum_{i=1}^{N}z_i^2\end{aligned}$$

Hence,

$$N^2\,(DO)^2=N^2m^2+N\sum_{i=1}^{N}z_i^2=N^2\,(AM)^2+N\sum_{i=1}^{N}z_i^2$$

Therefore, one obtains $DO\geq AM$ for multiple metrics cases. The equality sign only applies when all metric values are equal. **QED**.

In mathematics, the square-root of the average of the sum of squares is sometimes referred to as the quadratic mean (QM), i.e.,

$$QM=\sqrt{\frac{\sum_{i=1}^{N}m_i^2}{N}}\equiv DO.$$

For a set of positive real numbers, it is known [18], [19], [35] that $DO\geq AM\geq GM\geq HM$.

The situation with respect to DIP is explored below. In the 2-metric cases, $4\,(1-DIP)^2-4\,(1-AM)^2=2\left((1-x)^2+(1-y)^2\right)-4\left(1-\frac{x+y}{2}\right)^2=(x-y)^2\geq 0$. Thus, $(1-DIP)^2-(1-AM)^2\geq 0$. Taking the positive root, $AM\geq DIP$. Therefore, $DO\geq AM\geq DIP$.

Theorem 6: $AM\geq DIP$ is true in multiple metric cases.

Proof: As no proof of this theorem appears to exist in the literature, it is provided here. Now,

$$\begin{aligned}(1-DIP)^2&=\frac{1}{N}\sum_{i=1}^{N}(1-m_i)^2\\&=1-\frac{2}{N}\sum_{i=1}^{N}m_i+\frac{1}{N}\sum_{i=1}^{N}m_i^2\\&=1-2AM+DO^2\end{aligned}$$

In the multiple metric cases, it is already known that $DO\geq AM$. Therefore,

$$\begin{aligned}(1-DIP)^2&=1-2AM+DO^2\\&\geq 1-2AM+AM^2\\&=(1-AM)^2\end{aligned}$$

Taking the positive root, one obtains $AM\geq DIP$ for multiple metrics cases. The equality sign only applies when all metric values are equal. Therefore, $DO\geq AM\geq DIP$. **QED**.

Relations between DIP and GM as well as DIP and HM are more complicated. For example, there is a region of phase space where $GM<DIP$ and another region of phase space where $GM>DIP$. It can be shown that $GM=DIP$ on the curve described by $\sqrt{x}+\sqrt{y}=\sqrt{2}$. On the left-hand side of the curve (lower performance region) $GM<DIP$ and on the right-hand side of this curve (higher performance region) $GM>DIP$. Similarly, in the lower performance region of the phase space $HM<DIP$ and in the higher performance region of the phase space $HM>DIP$.

In summary, in the lowest performance region of the phase space $DO\geq AM\geq DIP\geq GM\geq HM$. In the lower performance region of the phase space $DO\geq AM\geq GM\geq DIP\geq HM$, while in the higher performance region of the phase space $DO\geq AM\geq GM\geq HM\geq DIP$.

TABLE 2. A toy example with two metrics (m_1, m_2) and five measures – *AM, GM. HM, DO,* and *DIP*.

Algorithms	m_1	m_2	*AM*	*GM*	*HM*	*DO*	*DIP*
A_1	0.59	**0.93**	**0.7600**	**0.7407**	**0.7220**	**0.7788**	*0.7059*
A_2	0.59	0.91	0.7500	0.7327	0.7159	0.7669	*0.7032*
A_3	**0.60**	0.90	0.7500	0.7348	0.7200	0.7649	***0.7085***

TABLE 3. Another toy example with two metrics (m_1, m_2) and five measures – *AM, GM. HM, DO,* and *DIP*.

Algorithms	m_1	m_2	*AM*	*GM*	*HM*	*DO*	*DIP*
A_1	0.8000	**0.8000**	0.8000	0.8000	***0.8000***	0.8000	***0.8000***
A_2	0.8293	0.7727	0.8010	0.8005	**0.8000**	0.8015	*0.7990*
A_3	**0.9000**	0.7200	**0.8100**	**0.8050**	**0.8000**	**0.8150**	*0.7898*

V. RELATIVE RANKING OF ALGORITHMS

In the Tables 2 to 7 below, each algorithm covers a row. A row consists of several cells; there are two cells for the two metric values (except for Table 7 which has three cells for three metric values) and five cells for *AM*, *GM*, *HM*, *DO*, and *DIP* measure values corresponding to the same pair (or trio) of metric values. To compare the performance of several algorithms, one can review only *AM* values or only *GM* values or only *HM* values or only *DO* values or only *DIP* values of these algorithms, as this is about relative ranking within one measure only. Therefore, the largest value within a column, representing a specific measure, corresponds to the best algorithm according to that specific measure.

TABLE 4. Based on published results on medical image segmentation using ISIC-2017 dataset [21] and two metrics (*Sensitivity* and *Specificity*).

Algor.	*Sensitivity*	*Specificity*	*AM*	*GM*	*HM*	*DO*	*DIP*
PSPNet [22]	0.7762	0.9782	0.8772	0.8714	0.8656	0.8830	*0.8410*
SESV-PSP [20]	0.8349	0.9561	0.8955	0.8934	0.8914	0.8975	*0.8792*
U-Net [23]	0.8098	**0.9826**	0.8962	0.8920	0.8879	0.9004	*0.8649*
SESV-Unet [20]	0.8440	0.9748	**0.9094**	**0.9070**	**0.9047**	**0.9117**	*0.8883*
FPN [24]	0.8094	0.9818	0.8956	0.8914	0.8873	0.8997	*0.8646*
SESV-FPN [20]	**0.8507**	0.9657	0.9082	0.9064	0.9046	0.9100	***0.8917***

TABLE 5. Based on published results on change detection of remote sensing images using LEVIR-CD dataset [28] and two metrics (*Recall* and *Precision*).

Algorithms	*Recall*	*Precision*	*AM*	*GM*	*HM*	*DO*	*DIP*
EC-EF [26]	0.9053	0.7496	0.8275	0.8238	0.8201	0.8311	*0.8107*
FC-Siam-Di [26]	**0.9292**	0.7818	0.8555	0.8523	0.8492	0.8587	*0.8378*
FC-Siam-Conc [26]	0.9163	0.7432	0.8297	0.8252	0.8207	0.8343	*0.8090*
FCN-PP [27]	0.8948	0.8031	0.8490	0.8477	0.8465	0.8502	*0.8421*
STANet [28]	0.8939	0.8614	0.8777	0.8775	0.8773	0.8778	***0.8766***
IFNet [29]	0.8652	**0.8755**	0.8703	0.8703	0.8703	0.8704	*0.8702*
FDCNN [30]	0.8871	0.8299	0.8585	0.8580	0.8575	0.8590	*0.8556*
SNUNet [32]	0.9134	0.8466	**0.8800**	**0.8794**	**0.8787**	**0.8806**	*0.8754*
DSAMNet [32]	0.8839	0.8275	0.8557	0.8552	0.8548	0.8562	*0.8530*

The objectives here are to find the best algorithm as well as to order the algorithms in the presence of several measures. In this case one needs to compare the *AM* value of an algorithm in a specific case with the *GM* value of the same algorithm in the same case and the *HM* value of the same algorithm in the case as well as the *DO* value of the same algorithm in the same case and the *DIP* value of the same algorithm in the same case.

For any pair of metrics (x, y), it has been proven in section IV-G that the value of $HM \leq$ the value of $GM \leq$ value of $AM \leq$ the value of DO. It has also been shown in section IV-C that the remaining phase space of *HM* under some specified constraint is less than the remaining phase space of *DIP*, while outside of that constraint the remaining phase space of *DIP* is less than the remaining phase space of *HM*. From considerations of both the nature of measure values and remaining phase spaces in conjunction with notions 1 (section II-B) and 2 (section IV-F), it is possible to write simple recipes for finding the best algorithm as well as the order of the algorithms in the presence of several measures in the following Tables.

1) It is clear from section IV-G that, independent of the region of the phase space, of these five measures either *HM* is the best or *DIP* is the best. In the lower performance region ($f < 0.63$) *HM* is the best and in the higher performance region ($f > 0.63$) *DIP* is the best for two-metric cases. Figure 5 offers a visual verification in the case of two metrics.
2) To find the best algorithm, find the smallest measure value along a row (i.e., for a specific algorithm). In each row (i.e., for each algorithm) there is one such smallest value (marked in green). Now find the largest of these smallest values in a Table. The algorithm corresponding to this largest value (marked in italic, green, and bold in the Tables below) is the best algorithm based on these measures and these metrics.
3) To order the algorithms, find the smallest measure value (marked in green in the Tables 2 to 7 below) along a row (i.e., for a specific algorithm). In each row (i.e., for each algorithm) there is one such smallest value (marked in green). Now order these smallest values from the largest to the smallest in a Table. The algorithm corresponding to the largest value is the best algorithm (marked in italic, green, and bold) and the algorithm corresponding to the smallest value in green is the worst algorithm based on these measures and these metrics. Basically, the orders of the algorithms follow the orders of smallest values in green.

VI. EXAMPLES

In this section, seven examples, both created ones and published ones, are reviewed to elucidate some points about these five measures. In each of Tables 2 to 7 the largest value within a column, representing a specific measure, corresponds to the best algorithm according to that specific measure and is highlighted in **bold**. The relative orders of the algorithms are highlighted in green and italic, with the cell containing green and bold number in italic corresponds to the best algorithm based on these measures and these metrics.

A. TABLE 2

Table 2 presents a toy example involving two metrics (m_1, m_2) and three algorithms A_1, A_2, A_3. According to m_1 and *DIP*, A_3 is the best, while A_1 is the best according to m_2, *AM*, *GM*, *HM*, and *DO*.

If one considers the ordering (best being the first) of the algorithms according to these metrics and measures, one finds $(A_3, A_1 = A_2)$ for m_1, (A_1, A_2, A_3) for m_2, $(A_1, A_2 = A_3)$ for *AM*, (A_1, A_3, A_2) for *GM*, (A_1, A_3, A_2) for *HM*, (A_1, A_2, A_3) for *DO*, and (A_3, A_1, A_2) for *DIP*. Of the five measures, *GM* and *HM* offer the same ordering while the remaining three differ from each other as well as from *GM* and *HM*. Following the above rules (see section V), it can be noted that

1) A_3 is the best algorithm. A_1 is the second-best algorithm. A_2 is the worst algorithm.
2) Only *DIP* offers the best ordering amongst the five measures.

B. TABLE 3

This is another toy example involving two metrics (m_1, m_2) and three algorithms A_1, A_2, A_3 presented in Table 3.

TABLE 6. Based on published results on change detection of remote sensing images using CCD dataset [33] and two metrics (*Recall* and *Precision*).

Algorithms	*Recall*	*Precision*	*AM*	*GM*	*HM*	*DO*	*DIP*
EC-EF [26]	0.8420	0.5267	0.6844	0.6659	0.6480	0.7023	*0.6472*
FC-Siam-Di [26]	0.7669	0.6185	0.6927	0.6887	0.6848	0.6967	*0.6839*
FC-Siam-Conc [26]	0.8044	0.4407	0.6225	0.5954	*0.5694*	0.6486	0.5810
FCN-PP [27]	0.9031	0.8169	0.8600	0.8589	0.8578	0.8611	*0.8535*
STANet [28]	0.9311	0.8898	0.9104	0.9102	0.9100	0.9107	*0.9081*
IFNet [29]	0.9176	0.8533	0.8854	0.8849	0.8843	0.8860	*0.8810*
FDCNN [30]	0.9170	0.8361	0.8765	0.8756	0.8747	0.8775	*0.8701*
SNUNet [31]	0.9475	0.9092	0.9284	0.9282	0.9280	0.9285	*0.9258*
DSAMNet [32]	**0.9483**	**0.9167**	**0.9325**	**0.9324**	**0.9322**	**0.9326**	***0.9307***

TABLE 7. Based on published results on medical image segmentation using ISIC-2017 dataset [21] and three metrics (*Accuracy*, *Sensitivity*, and *Specificity*).

Algorithms	Accuracy	*Sensitivity*	*Specificity*	*AM*	*GM*	*HM*	*DO*	*DIP*
PSPNet [22]	0.9221	0.7762	0.9782	0.8922	0.8880	0.8836	0.8962	*0.8626*
SESV-PSP [20]	0.9310	0.8349	0.9561	0.9073	0.9058	0.9042	0.9088	*0.8936*
U-Net [23]	0.9347	0.8098	**0.9826**	0.9090	0.9060	0.9030	0.9119	*0.8835*
SESV-Unet [20]	0.9417	0.8440	0.9748	**0.9202**	**0.9185**	**0.9167**	**0.9218**	*0.9028*
FPN [24]	0.9323	0.8094	0.9818	0.9078	0.9049	0.9018	0.9107	*0.8827*
SESV-FPN [20]	**0.9418**	**0.8507**	0.9657	0.9194	0.9180	0.9166	0.9207	***0.9054***

According to m_1, *AM*, *GM*, *HM*, and *DO*, A_3 is the best, while A_1 is the best according to m_2, *HM*, and *DIP*.

If one considers the ordering (best being the first) of the algorithms according to these metrics and measures, one finds (A_3, A_2, A_1) for m_1, (A_1, A_2, A_3) for m_2, (A_3, A_2, A_1) for *AM*, (A_3, A_2, A_1) for *GM*, $(A_1 = A_2 = A_3)$ for *HM*, (A_3, A_2, A_1) for *DO*, and (A_1, A_2, A_3) for *DIP*. Of the five measures, *AM*, *GM*, and *DO* offer the same ordering while *HM* cannot separate the three algorithms and *DIP* suggest a completely different ordering. Following the above rules (see section V), it can be noted that

1) A_1 is the best algorithm. A_2 is the second-best algorithm. A_3 is the worst algorithm.
2) Only *DIP* offers the best ordering amongst the five measures.

C. TABLE 4

In 2021 Xie et al. proposed the Segmentation-Emendation-reSegmentation-Verification (SESV) framework to improve the accuracy of existing medical image segmentation models [20]. They used, amongst others, the dataset provided by the International Skin Imaging Collaboration skin lesion segmentation challenges held in 2017 (ISIC-2017) [21]. They evaluated their SESV framework with PSPNet [22], U-Net [23], and FPN [24] as the base segmentation network and compared their results with the corresponding previously published results.

This is an example with real data, involving two metrics (*Sensitivity* and *Specificity*) and six algorithms. It can be observed that their framework in conjunction with X (where X is either PSPNet or UNet or FPN) improved the sensitivity compared with X in ISIC-2017 dataset but the specificity was reduced. Considering these two metrics in this dataset, which algorithm is the best?

Five measure have been calculated for each pair of sensitivity and specificity values for each of the aforementioned six algorithms and these are presented in Table 4. According to *AM*, *GM*, *HM*, and *DO*, SESV-UNet [20] is the best, while SESV-FPN [20] is the best according to *DIP*. All five measures agree that PPSNet [22] is the worst algorithm. Following the above rules (see section V), it can be noted that

1) SESV-FPN is the best algorithm and PSPNet is the worst algorithm.
2) Only *DIP* offers the best ordering amongst the five measures.

D. TABLE 5

Change detections in remote sensing images have many practical applications, e.g., environmental oversight, disaster monitoring, and urban planning. The aim of change detections is to identify differences between two images of the same geographical locations captured at two different times [25]. There are several state-of-the-art methods for change detection in remote sensing images, e.g., FC-EF [26], FC-Siam-Di [26], FC-Siam-Conc [26], FCN-PP [27], STANet [28], IFNet [29], FDCNN [30], SNUNet [31], and DSAMNet [32].

This is another example with real data. In this case the dataset is LEVIR-CD [28]. LEVIR-CD is a large publicly available change detection dataset with a variety of complex change features. In Table 5 there are two metrics (*Recall*, *Precision*) and nine algorithms. Five measures have been calculated for each pair of recall and precision values for each of these nine algorithms. According to *AM*, *GM*,

TABLE 8. Order of algorithms, based on two metrics (*Recall* and *Precision*) according to five measures (*AM, GM, HM, DO*, and *DIP*), extracted from the results in Table 5.

Best to worst algorithms	*AM*	*GM*	*HM*	*DO*	*DIP*
STANet [28]	SNUNet [32]	SNUNet [32]	SNUNet [32]	SNUNet [32]	*STANet [28]*
SNUNet [32]	STANet [28]	STANet [28]	STANet [28]	STANet [28]	*SNUNet [32]*
IFNet [29]	*IFNet [29]*	*IFNet [29]*	*IFNet [29]*	*IFNet [29]*	*IFNet [29]*
FDCNN [30]	*FDCNN [30]*	*FDCNN [30]*	*FDCNN [30]*	*FDCNN [30]*	*FDCNN [30]*
DSAMNet [32]	*DSAMNet [32]*	*DSAMNet [32]*	*DSAMNet [32]*	FC-Siam-Di [26]	*DSAMNet [32]*
FCN-PP [27]	FC-Siam-Di [26]	FC-Siam-Di [26]	FC-Siam-Di [26]	DSAMNet [32]	*FCN-PP [27]*
FC-Siam-Di [26]	FCN-PP [27]	FCN-PP [27]	FCN-PP [27]	FCN-PP [27]	*FC-Siam-Di [26]*
EC-EF [26]	FC-Siam-Conc [26]	FC-Siam-Conc [26]	FC-Siam-Conc [26]	FC-Siam-Conc [26]	*EC-EF [26]*
FC-Siam-Conc [26]	EC-EF [26]	EC-EF [26]	EC-EF [26]	EC-EF [26]	*FC-Siam-Conc [26]*
Algorithms in correct order	3 out of 9	3 out of 9	3 out of 9	2 out of 9	9 out of 9

HM, and *DO*, SNUNet [31] is the best, while STANet [28] is the best according to *DIP*. Considering this dataset, which algorithm is the best? Again, following the above rules (see section V), it is observed that, for the LEVIR-CD dataset,

1) STANet is the best algorithm and FC-Siam-Conc is the worst algorithm.
2) Only *DIP* offers the best ordering amongst the five measures.

E. TABLE 6

Similar to Table 5, Table 6 relates to change detections in remote sensing images. This is yet another example with real data. The dataset is CCD [33], which is also a publicly available dataset, capturing seasonal changes in the same geographical area from Google Earth. There are several state-of-the-art methods for change detection in remote sensing images, e.g., FC-EF [26], FC-Siam-Di [26], FC-Siam-Conc [26], FCN-PP [27], STANet [28], IFNet [29], FDCNN [30], SNUNet [31], and DSAMNet [32].

In Table 6 there are two metrics (*Recall*, *Precision*) and nine algorithms. Five measure have been calculated for each pair of recall and precision values for each of these nine algorithms. According to *AM*, *GM*, *HM*, *DO*, and *DIP*, DSMANet [32] is the best. Thus, all five measures concur on which algorithm is the best. Interestingly, in this case, all five measures concur that FC-Siam-Conc is the worst algorithm. Again, following the above rules (see section V), it is observed that, for the CCD dataset,

1) DSMASNet is the best algorithm and FC-Siam-Conc is the worst algorithm.
2) Amongst the five measures only *DIP* offers the best ordering of all algorithms.

F. TABLE 7

In this subsection the background research is the same as in section VI-C, involving the Segmentation-Emendation-reSegmentation-Verification (SESV) framework to improve the accuracy of existing medical image segmentation models [20]. Again, in this subsection, the real dataset was provided by the International Skin Imaging Collaboration skin lesion segmentation challenges held in 2017 (ISIC-2017) [21].

In subsection VI-C, only two metrics (*Sensitivity*, *Specificity*) and six algorithms have been considered. Now an additional metric (namely, *Accuracy*) is taken into account. Thus, the three metrics (*Accuracy*, *Sensitivity*, *Specificity*) and the same six algorithms are considered in this section and are presented in Table 7. It can be observed that their framework in conjunction with X (where X is either PSPNet or UNet or FPN) improved both the accuracy and sensitivity compared with X in ISIC-2017 dataset but the specificity was reduced. Considering these three metrics in this dataset, which algorithm is the best?

Five measure have been calculated for each tuple of accuracy, sensitivity and specificity values for each of these six algorithms and these are presented in Table 7. According to *AM*, *GM*, *HM*, and *DO*, SESV-UNet [20] is the best, while SESV-FPN [20] is the best according to *DIP*. All five measures agree that PPSNet [22] is the worst algorithm. Following the above rules (see section V), it can be noted that

1) SESV-FPN is the best algorithm and PSPNet is the worst algorithm.
2) Amongst the five measures only *DIP* offers the best ordering of all algorithms.

G. TABLE 8

This example with two metrics comes from the higher performance region. Summarising the results from Table 5, Table 8 presents, in columns 2 to 6, the order of algorithms, based on two metrics (*Recall* and *Precision*) according to five measures (*AM*, *GM*, *HM*, *DO*, and *DIP*). The column 1 presents the order of algorithms, considering the remaining phase space and following section V(3). The last row records the number of algorithms in the correct order, according to the column 1, for each of the five measures. In this case, of the nine algorithms *AM*, *GM*, *HM*, *DO*, and *DIP* correctly identify the order of 3, 3, 3, 2, and 9 algorithms respectively. Thus, *DO* finds the lowest number of algorithms in the correct order, while only *DIP* finds all nine algorithms in the correct order.

H. TABLE 9

This example with three metrics comes from the higher performance region. Summarising the results from Table 7, Table 9 presents, in columns 2 to 6, the order of algorithms,

TABLE 9. Order of algorithms, based on three metrics (*Accuracy, Sensitivity,* and *Specificity*) according to five measures (*AM, GM, HM, DO*, and *DIP*), extracted from the results in Table 7.

Best to worst algorithms	*AM*	*GM*	*HM*	*DO*	*DIP*
SESV-FPN [20]	SESV-Unet [20]	SESV-Unet [20]	SESV-Unet [20]	SESV-Unet [20]	*SESV-FPN [20]*
SESV-Unet [20]	SESV-FPN [20]	SESV-FPN [20]	SESV-FPN [20]	SESV-FPN [20]	*SESV-Unet [20]*
SESV-PSP [20]	U-Net [23]	U-Net [23]	*SESV-PSP [20]*	U-Net [23]	*SESV-PSP [20]*
U-Net [23]	FPN [24]	SESV-PSP [20]	*U-Net [23]*	FPN [24]	*U-Net [23]*
FPN [24]	SESV-PSP [20]	*FPN [24]*	*FPN [24]*	SESV-PSP [20]	*FPN [24]*
PSPNet [22]	*PSPNet [22]*	*PSPNet [22]*	*PSPNet [22]*	*PSPNet [22]*	*PSPNet [22]*
Algorithms in correct order	1 out of 6	2 out of 6	4 out of 6	1 out of 6	6 out of 6

based on three metrics (*Accuracy*, *Sensitivity*, and *Specificity*) according to five measures (*AM*, *GM*, *HM*, *DO*, and *DIP*). The column 1 presents the order of algorithms, considering the remaining phase space and following section V(3). The last row records the number of algorithms in the correct order, according to the column 1, for each of the five measures. In this case, of the six algorithms *AM*, *GM*, *HM*, *DO*, and *DIP* correctly identify 1, 2, 4, 1, and 6 respectively. Thus, *AM* and *DO* find the least number of algorithms in the correct order, while only *DIP* finds all six algorithms in the correct order.

VII. DISCUSSION

Below are some merited remarks:

1) In the case of a single metric, all the five measures (namely, *AM*, *GM*, *HM*, *DO*, and *DIP*) produce the identical outcome. This is not surprising, yet it is an important feature.
2) All the five measures (namely, *AM*, *GM*, *HM*, *DO*, and *DIP*), including the two proposed measures, are symmetric with respect to different metrics, which is essential if they are all independent and equally important.
3) In section IV, much of the explorations are in the case of two independent and equally important metrics to aid the visualisations of some selected example results. In this case, the phase space represents an area, but the phase space in any three metrics case will be a volume, while the phase space in more than three metrics case will be a hyper-volume.
4) In section IV-F, it is clear from Figures 4 and 5 that, for larger f values, the remaining phase space for *DIP* is smaller than any of the other four measures, i.e., *DIP* is the best. More than that is the fact that, even when f is asymptotically close to 1, *DIP* remains the best measure. Also, asymptotically as f tends to 1, all four ratios converge to the same value of $4/\pi$, implying that the performance of *AM*, *GM*, *HM*, and *DO* will be similar, even though it will remain the case that the performance of *DO* $<$ the performance of AM $<$ the performance of *GM* $<$ the performance of *HM*.
5) In the case of the metrics being independent but not equally important, one can extend these five measures in multiple-metric cases, using appropriate weightings of the metrics. For example, *HM* with two equally important metrics is popularly known as F_1 measure. A more general formulation is F_β measure [15], with unequal weights, can be written as
$$F_\beta = \left(\frac{1}{\beta^2+1}\frac{1}{m_1} + \left(1 - \frac{1}{\beta^2+1}\right)\frac{1}{m_2}\right)^{-1}$$
in the two-metric case. This can be further extended to multiple metric cases.
6) There are some similarities as well as differences amongst the five measures. But, *AM*, *GM*, and *HM* are not based on any relevant cost function except for some intuition. In contrast, the major advantage of the proposed *DO* and *DIP* is that they are based on explicit cost functions, which produce the correct result in the case of a single metric. Theoretical investigations in this paper demonstrate that *DO* is the worst of the five measures explored in this paper.
7) Further considerations of the cost functions as well as the remaining phase space, in the context of the problem, lead one to credit *DIP* as the better of the two proposed measures. It is not necessary to consider *AM*, *GM*, and *DO* for deciding which is the best algorithm, since either *HM* or *DIP* will be better in every region of the phase space.
8) Of these five measures, in two-metric cases the recommendation is to use *HM* for $f < 0.63$ (i.e., for the lower performance end) (see Figure 5). On the other hand, the recommendation is to use *DIP* for $f > 0.63$ (i.e., for the higher performance end) (see Figure 5).

VIII. CONCLUSION

To be able to perform comparative assessment of algorithms is essential in deciding the best algorithm and their rankings in data-driven decision making in any field. How to perform such an assessment objectively is obvious in the case of a single performance metric, but this is not so clear in the case of multiple metrics. In this paper, the harmonic mean (*HM*) [in two-metric cases, this is known as F_1 measure which is widely used], the arithmetic mean (*AM*), and the geometric mean (*GM*) have been reviewed. In the phase space of multiple-metric cases, it is always true that $DO \geq AM \geq GM \geq HM$, the equality is valid only when all metric values are equal.

The single metric case has been examined to develop two objective functions that are applicable for any number

of metrics. These two objective functions have led to two different performance measures – the distance from the origin (*DO*) and the distance from the ideal position (*DIP*). In the phase space of multiple-metric cases, it is proven that $DO \geq AM \geq DIP$, the equality is valid only when all metric values are equal.

A new concept of the remaining phase space for the evaluation of the quality of a performance measure is introduced in this paper. On further and closer examinations of the original goal and the remaining phase space of the metrics, amongst these five measures, either *HM* or *DIP* is the best. Moreover, it is proven that *HM* is the best measure at the lower performance end, but at the higher performance end, which is of much practical interest, *DIP* is clearly the best measure.

Rules for deciding the best algorithm and the order of a set of algorithms have been presented. Theoretical results have been derived in the context of multiple independent and bounded metrics. Furthermore, several properties of the five measures and detailed discussions have been provided. Some published comparison results have been reviewed in the present context to elucidate some points and to conclude that DIP is the best measure out of the five considered in the region of much practical interests.

ACKNOWLEDGMENT

The author would like to thank Dr C. Liu for formatting the manuscript.

REFERENCES

[1] B. Altinel and M. C. Ganiz, "Semantic text classification: A survey of past and recent advances," *Inf. Process. Manag.*, vol. 54, no. 6, pp. 1129–1153, Nov. 2018.

[2] Z. Li, Y. Fan, B. Jiang, T. Lei, and W. Liu, "A survey on sentiment analysis and opinion mining for social multimedia," *Multimedia Tools Appl.*, vol. 78, no. 6, pp. 6939–6967, 2019.

[3] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *J. Intell. Learn. Syst. Appl.*, vol. 9, no. 1, pp. 1–16, 2017.

[4] J. L. Martinez-Rodriguez, A. Hogan, and I. Lopez-Arevalo, "Information extraction meets the semantic web: A survey," *Semantic Web*, vol. 11, no. 2, pp. 255–335, Feb. 2020.

[5] M. Zhu, Z. Chen, and Y. Yuan, "DSI-Net: Deep synergistic interaction network for joint classification and segmentation with endoscope images," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3315–3325, Dec. 2021, doi: 10.1109/TMI.2021.3083586.

[6] S. Mishra, Y. Zhang, D. Z. Chen, and X. S. Hu, "Data-driven deep supervision for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 6, pp. 1560–1574, Jun. 2022, doi: 10.1109/TMI.2022.3143371.

[7] Z. Liao, Y. Xie, S. Hu, and Y. Xia, "Learning from ambiguous labels for lung nodule malignancy prediction," *IEEE Trans. Med. Imag.*, vol. 41, no. 7, pp. 1874–1884, Jul. 2022, doi: 10.1109/TMI.2022.3149344.

[8] T. Xiang, Y. Song, C. Zhang, D. Liu, M. Chen, F. Zhang, H. Huang, L. O'Donnell, and W. Cai, "DSNet: A dual-stream framework for weakly-supervised gigapixel pathology image analysis," *IEEE Trans. Med. Imag.*, vol. 41, no. 8, pp. 2180–2190, Aug. 2022, doi: 10.1109/TMI.2022.3157983.

[9] Y. Xu, Y. Yang, E. Wang, F. Zhuang, and H. Xiong, "Detect professional malicious user with metric learning in recommender systems," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 9, pp. 4133–4146, Sep. 2022, doi: 10.1109/TKDE.2020.3040618.

[10] T. Lei, D. Xue, H. Ning, S. Yang, Z. Lv, and A. K. Nandi, "Local and global feature learning with kernel scale-adaptive attention network for VHR remote sensing change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7308–7322, 2022, doi: 10.1109/JSTARS.2022.3200997.

[11] W. Wu, L. He, W. Lin, Y. Su, Y. Cui, C. Maple, and S. Jarvis, "Developing an unsupervised real-time anomaly detection scheme for time series with multi-seasonality," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 9, pp. 4147–4160, Sep. 2022, doi: 10.1109/TKDE.2020.3035685.

[12] J.-G. Choi, I. Ko, J. Kim, Y. Jeon, and S. Han, "Machine learning framework for multi-level classification of company revenue," *IEEE Access*, vol. 9, pp. 96739–96750, 2021, doi: 10.1109/ACCESS.2021.3088874.

[13] S. A. Salem, N. M. Salem, and A. K. Nandi, "Segmentation of retinal blood vessels using a novel clustering algorithm (RACAL) with a partial supervision strategy," *Med. Biol. Eng. Comput.*, vol. 45, no. 3, pp. 261–273, Feb. 2007, doi: 10.1007/s11517-006-0141-2.

[14] J. Li, C. Wang, J. Chen, H. Zhang, Y. Dai, L. Wang, L. Wang, and A. K. Nandi, "Explainable CNN with fuzzy tree regularization for respiratory sound analysis," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 6, pp. 1516–1527, Jun. 2022, doi: 10.1109/TFUZZ.2022.3144448.

[15] N. Chinchor, "MUC-4 evaluation metrics," in *Proc. 4th Conf. Message Understand. (MUC)*, 1992, pp. 22–29. [Online]. Available: https://aclanthology.org/M92-1002.pdf

[16] Y. Sasaki. (2007). *The Truth of F-Measure*. Accessed: Sep. 7, 2022. [Online]. Available: https://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf

[17] C. J. V. Tijsbergen. *Information Retrieval*. Accessed: Sep. 7, 2022. [Online]. Available: http://www.dcs.gla.ac.uk/Keith/Preface.html

[18] *Inequality of Arithmetic and Geometric Means*. Accessed: Sep. 7, 2022. [Online]. Available: https://en.wikipedia.org/wiki/Inequality_of_arithmetic_and_geometric_means#:~:text=equal%2C%20as%20desired.-,The%20case%20where%20not%20all%20the%20terms%20are%20equal,greater%20than%20the%20geometric%20mean

[19] *HM-GM-AM-QM Inequalities*. Accessed: Sep. 7, 2022. [Online]. Available: https://en.wikipedia.org/wiki/HM-GM-AM-QM_inequalities

[20] Y. Xie, J. Zhang, H. Lu, C. Shen, and Y. Xia, "SESV: Accurate medical image segmentation by predicting and correcting errors," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 286–296, Jan. 2021, doi: 10.1109/TMI.2020.3025308.

[21] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Oct. 2018, pp. 168–172.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent*, 2015, pp. 234–241.

[24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, Jul. 2017, pp. 936–944.

[25] A. Sebastian, T. Tuma, N. Papandreou, M. L. Gallo, L. Kull, T. Parnell, and E. Eleftheriou, "Temporal correlation detection using computational phase-change memory," *Nature Commun.*, vol. 8, no. 1, pp. 1–10, 2017, doi: 10.1038/s41467-017-01481-9.

[26] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4063–4067.

[27] T. Lei, Y. Zhang, Z. Lv, S. Li, S. Liu, and A. K. Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 982–986, Jun. 2019, doi: 10.1109/LGRS.2018.2889307.

[28] H. Chen and Z. Shi, "A spatial–temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, 2020, doi: 10.3390/rs12101662.

[29] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution Bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020.

[30] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, 2020, doi: 10.1109/TGRS.2020.2981051.
[31] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, Art. no. 8007805, doi: 10.1109/LGRS.2021.3056416.
[32] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604816, doi: 10.1109/TGRS.2021.3085870.
[33] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 42, no. 2, pp. 565–571, Jun. 2018.
[34] C. J. V. Tijsbergen, *Information Retrieval*. London, U.K.: Butterworths, 1979.
[35] H. Sedrakyan and N. Sedrakyan, *Algebraic Inequalities*, Berlin, Germany: Springer, 2018.

ASOKE K. NANDI (Life Fellow, IEEE) received the Ph.D. degree in physics from the University of Cambridge (Trinity College), Cambridge.

He held academic positions in several universities, including Oxford, Imperial College London, Strathclyde, and Liverpool, and Finland Distinguished Professorship in Jyvaskyla. In 2013, he moved to Brunel University London, to become the Chair and the Head of Electronic and Computer Engineering. In 1983, he co-discovered the three fundamental particles known as W^+, W^-, and Z^0 (by the UA1 team at CERN), providing the evidence for the unification of the electromagnetic and weak forces, for which the Nobel Committee for Physics awarded the prize to his two team leaders for their decisive contributions, in 1984. His current research interests include signal processing and machine learning, with applications to communications, image segmentations, and biomedical data. He has made many fundamental theoretical and algorithmic contributions to many aspects of signal processing and machine learning. He has authored over 600 technical publications, including 280 journal articles and six books, entitled *Image Segmentation: Principles, Techniques, and Applications* (Wiley, 2022), *Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines* (Wiley, 2020), *Integrative Cluster Analysis in Bioinformatics* (Wiley, 2015), *Automatic Modulation Classification: Principles, Algorithms and Applications* (Wiley, 2015), *Blind Estimation Using Higher-Order Statistics* (Springer, 1999), and *Automatic Modulation Recognition of Communications Signals* (Springer, 1996). The H-index of his publications is 80 (Google Scholar) and his ERDOS number is 2.

Professor Nandi is a Fellow of the Royal Academy of Engineering (U.K.) and a Fellow of seven other institutions, including the IEEE and the IET. Among the many awards, he received are the Institute of Electrical and Electronics Engineers (USA) Heinrich Hertz Award, in 2012, the Glory of Bengal Award for his outstanding achievements in scientific research, in 2010, the Water Arbitration Prize of the Institution of Mechanical Engineers, U.K., in 1999, and the Mountbatten Premium, Division Award of the Electronics and Communications Division, Institution of Electrical Engineers, U.K., in 1998. He was an IEEE EMBS Distinguished Lecturer (2018–2019).