IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 22, 2019, accepted November 7, 2019, date of publication November 12, 2019,
date of current version November 21, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2953104

A Simplified Cohen'S Kappa for Use in Binary Classification Data Annotation Tasks

JUAN WANG[1], (Member, IEEE), YONGYI YANG[2], AND BIN XIA[3]
[1]Delta Micro Technology Inc., Laguna Hills, CA 92603, USA
[2]Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616, USA
[3]Shenzhen SiBright Company Ltd., Shenzhen 518000, China

Corresponding author: Juan Wang (wangjuan313@gmail.com)

ABSTRACT In binary classification tasks, Cohen's kappa is often used as a quality measure for data annotations, which is inconsistent with its original purpose as an inter-annotator consistency measure. The analytic relationship between kappa and commonly used classification metrics (e.g., sensitivity and specificity) is nonlinear, and thus is difficult to be applied for interpretation of the classification performance (merely from the knowledge of the kappa value) of the annotations. In this study, based on an annotation generation model, we derive a simplified, linear relationship for Cohen's kappa, sensitivity, and specificity by using the 1st-order Taylor approximation. This relationship is further simplified by relating to Youden's J statistic, a performance metric for binary classification tasks. We provide an analysis on the linear coefficients in the simplified relationship and the approximation error, and conduct a linear regression analysis to assess the relationship by using a synthetic dataset where the ground truth is known. The results show that there is only a negligible approximation error in the simplified relationship when no major bias and prevalence issues exist. Furthermore, the relationship between kappa and Youden's J is validated on an annotation dataset from seven graders in a diabetic retinopathy screening study. The discrepancy between kappa and Youden's J is demonstrated to be an effective measure for annotator assessment when no ground truth is available.

INDEX TERMS Cohen's kappa, sensitivity, specificity, Youden's J statistic, relationship, annotator evaluation.

I. INTRODUCTION

In medical imaging, almost all state-of-the-art methods for lesion detection and disease diagnosis tasks are developed by applying supervised learning with a binary classification formulation [1]–[4]. For this purpose, annotations of the training samples (i.e. labels) need to be obtained beforehand. In the machine learning community, it has been demonstrated that the quality of the data labels can have a number of effects on the resulting classifier, ranging from the classification performance, the complexity of the classifier model, to the number of required training samples [5]–[9]. For example, Pelletier *et al.* [10] studied the effect of annotation noise (i.e., errors) on classification performance in land cover mapping from satellite time series images; it was concluded that the classifier performance can be adversely affected when the noise levels are higher than 25%–30%.

While in practice annotations are typically obtained from experts for the annotation task under consideration, it is often difficult, if not impossible, to achieve perfect annotations [11], [12]. This can be due to the subjectivity of the annotators, the imperfect knowledge of the experts, or the difficulty of the annotation tasks. This is especially the case in biomedical data [13]–[15], in which the annotation tasks are often difficult and complex. Therefore, it is important to assess the quality of annotations prior to their use in supervised learning.

In the literature, a common approach to measure the quality of annotations is to apply a metric to assess the inter-annotator consistency in the data [16]. One such metric is Cohen's kappa coefficient (or kappa in short) [17], which has been accepted as the de facto standard for measurement of inter-annotator agreement [18], [19]. Mathematically, Cohen's kappa is defined as:

$$\kappa = \frac{p_A - p_E}{1 - p_E} \tag{1}$$

where p_A is the observed relative agreement between two annotators, and p_E is the hypothetical probability of agreement by chance (with data labels randomly assigned).

The associate editor coordinating the review of this manuscript and approving it for publication was Mingjun Dai.

TABLE 1. Two example sets of diagnostic results compared with the gold standard. Plus signs indicate the positive class, while negative signs indicate the negative class.

		Gold standard		
		+	-	Total
Diagnosis	+	81	1	82
	-	9	9	18
	Total	90	10	100

(a)

Gold standard		
+	-	Total
93	2	95
0	5	5
93	7	100

(b)

In particular, $\kappa = 1$ corresponds to the case of perfect agreement, whereas $\kappa = 0$ indicates no agreement other than what would be expected by chance. While commonly used, Cohen's kappa is also cited for its problems associated with bias and prevalence in the interpretation of kappa values [18], [20]. The bias problem is caused by the difference in the distribution of annotation categories of the two annotators, while the prevalence problem arises when the underlying distribution of class categories is skewed [18].

It is noted that Cohen's kappa is only intended to evaluate how often the annotators may agree with each other. It does not, however, measure directly the quality (i.e., the accuracy) of annotations for a classification task, where sensitivity and specificity are at the most concern. This is illustrated by examples shown in Table 1, wherein the confusion matrices of two diagnostic experiments are given. In the table, plus signs denote the positive class and negative signs the negative class. In Table 1(a), both the sensitivity and specificity of the diagnostic results are at 90%, a very high performance for a binary classification task; however, its kappa value is only 0.590, indicating only a moderate agreement between the diagnostic results and the gold standard. On the other hand, in Table 1(b), the sensitivity and specificity of the diagnostic results are at 100% and 71.4%, respectively, a relatively low specificity; however, the kappa value is 0.823, indicating a high agreement between the diagnostic results and the gold standard.

Without the knowledge on the accuracy of annotations in terms of classification performance, it is difficult to assess their efficacy for classifier training. In the literature, the threshold of acceptability is often set empirically based on the kappa value. For example, kappa value of 0.67 is used as a cutoff in computational linguistics [18], [21]. However, there is little understanding of how well the annotations are merely judging from the kappa value. As illustrated in Table 1 above, there is a noticeable gap between the use of kappa as an inter-annotator consistency measure and its use as a quality measure of annotations in a binary classification task.

Because of this noted inconsistency in the use of kappa, there have been great interests in investigating the relationship between Cohen's kappa and the performance metrics commonly used in classification tasks (i.e. sensitivity and specificity) [22]–[25]. Several studies have derived independently an analytic relationship of kappa, sensitivity, and specificity [22], [23]. For example, based on the 2×2 confusion matrix, Feuerman and Miller [23] obtained the following relationship:

$$\kappa = \frac{2\alpha\beta(Se + Sp - 1)}{(\alpha^2 + \beta^2) + (\beta - \alpha)(\alpha Se - \beta Sp)}$$

for $Se \neq Sp$, where Se is sensitivity, Sp is specificity, α is the proportion of examples in the positive class, and $\beta = 1 - \alpha$. However, this relationship is often difficult to be employed for interpretation, because of the nonlinear nature of sensitivity and specificity in the function.

In a previous study, based on an annotation generation model, we derived a linear relationship $\kappa = Se + Sp - 1$ for unbiased annotations [26]. In this study we extend this derivation to the more general case of biased annotations. We derive a simplified, linear relationship of kappa, sensitivity and specificity by employing the 1st-order Taylor approximation. This relationship is further simplified by relating to Youden's J statistic, a metric used for classification performance. To help elucidating this relationship, we provide an analysis on the linear coefficients and validate the approximation error empirically. For the latter, a linear regression analysis is performed to compare the true relationship with the simplified one based on a synthetic dataset. The results demonstrate the effectiveness of the developed relationship when no severe bias and prevalence issues exist. In addition, the relationship between kappa and Youden's J is also validated on a real-life dataset collected from a diabetic retinopathy (DR) screening study, wherein the discrepancy between kappa and Youden's J is applied for annotator assessment.

II. METHODS

A. ANNOTATION GENERATION MODEL

In an annotation task, the instances under consideration can be thought as being drawn from a population that is a mixture of two subpopulations [27], [28]: reliable and unreliable. The reliable subpopulation consists of instances that are easy to annotate, so that the two annotators will always agree on their labels [27], [28]. The unreliable subpopulation consists of instances that are hard to annotate, so that the two annotators will be agreed on their labels by chance alone [27], [28].

Let X_i be the annotation of an instance provided by the ith annotator ($i \in \{1, 2\}$) and $c \in \{0, 1\}$ be the category of the labels, in which $c = 1$ denotes positive class and $c = 0$ is negative class. The annotation process above suggests an annotation generation model with latent variable l for the easy and hard types, i.e., $l = E$ (easy) and $l = H$ (hard) [28]. It is described by conditions as follows:

$$P(X_1 = c|l = E) = P(X_2 = c|l = E) \quad (2)$$

$$P(X_1 = X_2|l = E) = 1 \quad (3)$$

$$P(X_1, X_2|l = H) = P(X_1|l = H)P(X_2|l = H) \quad (4)$$

Equations (2) and (3) represents that the two annotators perfectly agree on the easy instances. Equation (4) denotes that the two annotators independently provide labels for the hard instances. Note when one annotator is ground-truth annotator,

TABLE 2. Mathematical symbols used in this study.

symbols	expressions	properties
p_c	$P(X_1 = c \mid l = H)$	$p_0 + p_1 = 1$
q_c	$P(X_2 = c \mid l = H)$	$q_0 + q_1 = 1$
e	$P(l = E)$	$e + h = 1$
h	$P(l = H)$	
p_e	$P(X_1 = X_2 \mid l = E)$	$p_e = 1$
p_h	$P(X_1 = X_2 \mid l = H)$	$p_h = p_0 q_0 + p_1 q_1$

equation (4) still holds. In this case, the other annotator makes guessing on the hard instances, thus the resulting labels are independent of their true labels (obtained from the ground-truth annotator).

Based on the annotation generation model, for ease of development, the mathematical symbols of different expressions used in this study are listed in Table 2.

From equations (3) and (4), the probabilities of the two annotators agree on the easy and hard instances are:

$$p_e = P(X_1 = X_2 \mid l = E) = 1 \tag{5}$$

and

$$\begin{aligned} p_h &= P(X_1 = X_2 \mid l = H) \\ &= P(X_1 = 0 \mid l = H) P(X_2 = 0 \mid l = H) \\ &\quad + P(X_1 = 1 \mid l = H) P(X_2 = 1 \mid l = H) \\ &= p_0 q_0 + p_1 q_1 \end{aligned} \tag{6}$$

respectively.

The annotation generation model yields $\kappa = 1$ for easy instances and $\kappa = 0$ for hard instances [26]. When the population consists of both easy and hard instances, $0 < \kappa < 1$. These results together indicate that $0 \leq \kappa \leq 1$ when the annotation generation model is considered. Note the definition in equation (1) indicates that it is possible for kappa to be negative [29], which implies that the agreement of the two annotators is worse than random. Negative kappa happens rarely in real annotation tasks, thus is beyond the scope of this paper.

B. COHEN'S KAPPA COEFFICIENT

To investigate the relationship between kappa and classification metrics, one annotator has to be ground-truth annotator. Without loss of generality, let X_2 be the ground-truth annotator. Assume distributions of different categories in easy and hard subpopulations are same for ground truths, i.e., $P(X_2 = c \mid l = E) = P(X_2 = c \mid l = H) = q_c$, the proportion of category c in the population is:

$$P(X_2 = c) = q_c \tag{7}$$

Note $q_1 = 0$ (i.e. $q_0 = 1$) and $q_0 = 0$ (i.e. $q_1 = 1$) will lead to undefined sensitivity and specificity, respectively. Therefore, this study only considers $0 < q_c < 1$ (i.e. both categories are present in the population).

With the above assumption, it can be easily shown that $P(l \mid X_2 = 0) = P(l \mid X_2 = 1) = P(l)$, suggesting that proportions of hard (and easy) instances in category 0 is equal to that in category 1. It is approximately true in most applications considering annotators always have difficulty in discriminating the boundary instances, which validates the effectiveness of above assumption in real application.

Moreover, equation (2) yields $P(X_1 = c \mid l = E) = P(X_2 = c \mid l = E) = q_c$, thus the proportion of instances in category c for X_1 is:

$$P(X_1 = c) = q_c e + p_c h \tag{8}$$

From equations (7) and (8), the chance agreement and the relative observed agreement between two annotators are as follows:

$$p_E = (1 - 2q_0 q_1) e + p_h h \tag{9}$$

and

$$p_A = e + p_h h \tag{10}$$

In the end, Cohen's kappa can be expressed as:

$$\kappa = \frac{e}{e + \frac{1 - p_h}{2 q_0 q_1} h} \tag{11}$$

with $0 \leq \kappa \leq 1$. For conciseness, the derivations of kappa (including equations (7), (8), (9), (10) and (11)) and its range are shown in Appendix A.

C. SENSITIVITY AND SPECIFICITY

The probability that the two annotators agree on category c is obtained as follows:

$$P(X_1 = c, X_2 = c) = q_c e + p_c q_c h \tag{12}$$

The derivation of which is shown in Appendix B.

From equations (7) and (12), sensitivity and specificity of X_1 can be calculated as:

$$\begin{aligned} Se &= P(X_1 = 1 \mid X_2 = 1) \\ &= \frac{P(X_1 = 1, X_2 = 1)}{P(X_2 = 1)} \\ &= p_0 e + p_1 \end{aligned} \tag{13}$$

and

$$\begin{aligned} Sp &= P(X_1 = 0 \mid X_2 = 0) \\ &= \frac{P(X_1 = 0, X_2 = 0)}{P(X_2 = 0)} \\ &= p_1 e + p_0 \end{aligned} \tag{14}$$

respectively, which further yield the relationship as follows:

$$Se + Sp = 1 + e \tag{15}$$

This relationship indicates that the summation of sensitivity and specificity is determined by the proportion of easy instances in the dataset.

Youden's J statistic [30], [31] is a performance summary for binary classification task. It is defined as $J \triangleq Se + Sp - 1$.

Equation (15) suggests the following relationship between e and Youden's J:

$$J = e \tag{16}$$

It indicates that Youden's J measures the proportion of the easy instances.

D. KAPPA APPROXIMATION

The kappa expression in equation (11) is complex and difficult to be interpreted by sensitivity and specificity. To deal with this issue, we consider kappa approximation in this section. To ensure low approximation error, we derive kappa approximation with respect to e and h, respectively.

For simplicity of notations, let $B = \frac{1-p_h}{2q_0q_1}$, then

$$B = \frac{1-p_h}{2q_0q_1} = \frac{1}{2}\left(\frac{p_0}{q_0} + \frac{1-p_0}{1-q_0}\right) > \frac{1}{2} \tag{17}$$

1) KAPPA APPROXIMATION WITH RESPECT TO E

For $e > 0$, equation (11) can be formulated as:

$$\kappa = \frac{1}{1+B(1/e-1)} = \frac{1}{1-B}\left(1 - \frac{1}{1-\frac{B-1}{B}e}\right) \tag{18}$$

Recall the Taylor series of $\frac{1}{1-x}$ for $|x| < 1$ is:

$$\frac{1}{1-x} = \sum_{i=0}^{\infty} x^i \tag{19}$$

Then kappa in equation (18) is expressed as:

$$\kappa = \frac{e}{B}\sum_{i=1}^{\infty}\left(\frac{B-1}{B}e\right)^{i-1} \tag{20}$$

Note $|\frac{B-1}{B}e| < 1$ holds for $0 < q_0 < 1$, which is proved in Appendix C.

Therefore, for $e > 0$, κ is approximated as:

$$\kappa \approx e/B \tag{21}$$

according to the 1st-degree Taylor expansion. Moreover, from equation (11), $\kappa = 0$ for $e = 0$, thus $\kappa = e/B$ holds as well.

The use of 1st-order Taylor expansion in equation (21) unavoidably introduces error as follows:

$$error = \frac{e}{B}\sum_{i=2}^{\infty}\left(\frac{B-1}{B}e\right)^{i-1} \tag{22}$$

When B is fixed, the error is monotonically increasing with respect to e. Therefore, the above approximation may introduce great error when e is large.

2) KAPPA APPROXIMATION WITH RESPECT TO H

To eliminate the issue of potential great error when e is large, we further introduce a kappa approximation with respect to h, for which equation (11) is formulated as:

$$\kappa = \frac{1}{1+Bh/(1-h)} = \frac{1}{1-B}\left(1 - \frac{B}{1-(1-B)h}\right) \tag{23}$$

Based on Taylor series, kappa in equation (23) can be expressed as:

$$\kappa = 1 - Bh\sum_{i=1}^{\infty}[(1-B)h]^{i-1} \tag{24}$$

Considering the 1st-degree Taylor expansion, κ is approximated as:

$$\kappa \approx 1 - Bh \tag{25}$$

which leads to error as follows:

$$error = Bh\sum_{i=2}^{\infty}[(1-B)h]^{i-1} \tag{26}$$

3) FINAL KAPPA APPROXIMATION

From equations (21) and (25), kappa approximation with respect to e and h yield errors:

$$error_e = \frac{e}{e+Bh} - e/B = \frac{e^2(B-1)}{B(e+Bh)} \tag{27}$$

and

$$error_h = \frac{e}{e+Bh} - (1-Bh) = \frac{h^2B(B-1)}{e+Bh} \tag{28}$$

respectively.

Solving $|error_e| \le |error_h|$ gets $e \le \frac{B}{1+B}$. Therefore, the final kappa approximation is:

$$\kappa \approx \begin{cases} e/B, & \text{if } e \le \dfrac{B}{1+B} \\ Be + 1 - B, & \text{otherwise} \end{cases} \tag{29}$$

in which kappa approximation with respect to e is in favor of low e, while kappa approximation with respect to h is in favor of low h (thus high e).

E. RELATIONSHIP OF KAPPA, SENSITIVITY AND SPECIFICITY

From equations (15), (16), and (29), the relationship of kappa, sensitivity, and specificity is obtained as follows:

$$\begin{aligned} J &= Se + Sp - 1 \\ &\approx \begin{cases} B\kappa, & \text{if } e \le \dfrac{B}{1+B} \\ \dfrac{1}{B}\kappa + \dfrac{B-1}{B}, & \text{otherwise} \end{cases} \end{aligned} \tag{30}$$

Since $B > 0.5$, this relationship indicates that for fixed B, as κ increases, J increases. More importantly, equation (30) indicates some undesirable relationships between kappa and Youden's J for small and large B's as follows, which should be avoided in real application: 1) when B is high, increasing value of κ increases J dramatically for $e \le \frac{B}{1+B}$, but has very little effect on the value of J for $e > \frac{B}{1+B}$; 2) when B is low, increasing value of κ has very little effect on the value of J for $e \le \frac{B}{1+B}$ but increases J dramatically for $e > \frac{B}{1+B}$.

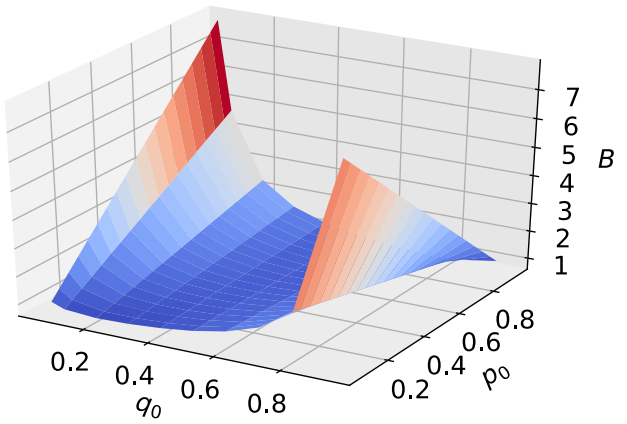

FIGURE 1. 3D plot of function $B(p_0, q_0)$.

For the special case of unbiased annotations (i.e. $p_c = q_c$), we get $B = 1$ from equation (17) and the relationship in equation (30) can be simplified as follows:

$$J = Se + Sp - 1 = \kappa = e \tag{31}$$

This relationship has been developed in [26] as well.

III. FURTHER ANALYSIS

A. ANALYSIS OF B

Due to the importance of B in the relationship (30), in this section we will analyze the value of B in different aspects of views as follows: 1) the range of B, 2) the effect of (q_0, p_0) on B, and 3) the effect of prevalence on B.

1) RANGE OF B

Equation (17) has been shown that $B > 0.5$. In particular, $B \to \infty$ when $q_0 \to 0$ or $q_0 \to 1$ (i.e. extremely severe prevalence problem is present). In this study, we consider less severe prevalence problems with $0.06 \le q_0, p_0 \le 0.94$ and Figure 1 shows the corresponding 3D plot of function $B(q_0, p_0)$. Epirically, the range of B is $0.74 \le B \le 7.87$ for $0.06 \le q_0, p_0 \le 0.94$.

More importantly, from Figure 1, it can be seen that B has relatively small values for most (q_0, p_0)'s, and high values when 1) $q_0 \to 0$ and $p_0 \to 1$, or 2) $q_0 \to 1$ and $p_0 \to 0$. These observations imply that extremely large B occurs when the so-called prevalence (i.e. q_0 is far from 0.5) and bias (i.e. p_0 is far from q_0) problems are present [18], [20]. The prevalence problem should be avoided by selecting relatively balanced dataset for annotation; the bias problem should be avoided by annotator selection and annotation method design, which have been investigated in our previous study for diabetic retinopathy grading in fundus images [32].

2) EFFECT OF (q_0, p_0) ON B

As noted in Section II-E, $B = 1$ corresponds to unbiased annotations. Therefore, good annotations get $B \to 1$, which also yields $B \to 1/B$. For simplicity, this study considers the effect of (q_0, p_0) for $0.5 < B < 2$, which also yields $0.5 < 1/B < 2$. With simple mathematical calculation, the solutions of $0.5 < B < 2$ are obtained as follows:

1) $0 < q_0 < \frac{1}{4}, 0 < p_0 < \frac{3q_0 - 4q_0^2}{1-2q_0}$;
2) $\frac{1}{4} \le q_0 \le \frac{3}{4}, 0 < p_0 < 1$; or
3) $\frac{3}{4} < q_0 < 1, 1 > p_0 > \frac{3q_0 - 4q_0^2}{1-2q_0}$.

These solutions demonstrate that majority of (q_0, p_0) satisfies $0.5 < B < 2$.

From the above solutions, it can be seen that when $0 < q_0 < \frac{1}{4}$, $\frac{3q_0 - 4q_0^2}{1-2q_0}$ decreases as q_0 decreases (i.e. more severe prevalence problem), thus the range of p_0 decreases; similarly, when $\frac{3}{4} < q_0 < 1$, $\frac{3q_0 - 4q_0^2}{1-2q_0}$ increases when q_0 increases (i.e. more severe prevalence problem), thus the range of p_0 decreases as well. These results suggest that starting from prevalence index $|q_0 - 0.5| = 0.25$, the range of p_0 satisfying $0.5 < B < 2$ become narrower when the prevalence problem becomes severe (i.e. $|q_0 - 0.5|$ becomes larger). These results imply that for $B \to 1$ in a fixed small range, the tolerance to bias problem becomes less when the prevalence problem becomes severe.

3) EFFECT OF PREVALENCE ON B

Let $x = q_0 - 0.5$, $-0.5 < x < 0.5$, then

$$B(x, p_0) = \frac{p_0}{0.5 + x} + \frac{1 - p_0}{0.5 - x} \tag{32}$$

which yields

$$\frac{\partial B(x, p_0)}{\partial x} = \frac{(1 - 2p_0)x^2 + x + 0.25(1 - 2p_0)}{(0.25 - x^2)^2} \tag{33}$$

Since $(0.25 - x^2)^2 > 0$, the sign of $\frac{\partial B(x,p_0)}{\partial x}$ is determined by its nominator $(1 - 2p_0)x^2 + x + 0.25(1 - 2p_0)$. Let $f(x) = (1 - 2p_0)x^2 + x + 0.25(1 - 2p_0)$, $f(x)$ is a quadratic function respect to x. Its two solutions are $x_1 = \frac{-0.5 - \sqrt{p_0 - p_0^2}}{1 - 2p_0}$ and $x_2 = \frac{-0.5 + \sqrt{p_0 - p_0^2}}{1 - 2p_0}$, and its local maximum or minimum is obtained at $x_0 = \frac{1}{2(2p_0 - 1)}$ with $f(x_0) = \frac{p_0(1-p_0)}{2p_0 - 1}$. Furthermore, we can get $f(-0.5) = -p_0$ and $f(0.5) = 1 - p_0$. With the above facts, it can be easily proved that:

i) for a fixed $p_0 \in (0, 0.5)$, $B(x, p_0)$ is a decrease function respect to x for $-0.5 < x < x_2$, an increase function respect to x for $x_2 < x < 0.5$.

ii) for a fixed $p_0 \in (0.5, 1)$, $B(x, p_0)$ is a decrease function respect to x for $-0.5 < x < x_1$, an increase function respect to x for $x_1 < x < 0.5$.

The above conclusions indicate that $B(x, p_0)$ has its minimum at x_2 for a fixed $p_0 \in (0, 0.5)$ and at x_1 for a fixed $p_0 \in (0.5, 1)$. For a fixed p_0, the value of $B(x, p_0)$ tends to increase as the prevalence problem becomes severe (i.e. $|x| \to 0.5$).

TABLE 3. Interpretation of kappa value.

kappa value	agreement
$\kappa = 0$	change agreement
$0 < \kappa \leq 0.2$	slight agreement
$0.2 < \kappa \leq 0.4$	fair agreement
$0.4 < \kappa \leq 0.6$	moderate agreement
$0.6 < \kappa \leq 0.8$	substantial agreement
$0.8 < \kappa < 1$	almost perfect agreement
$\kappa = 1$	perfect agreement

B. ERROR ANALYSIS IN KAPPA APPROXIMATION

From equations (27) and (28), the kappa approximation error can be calculated as follows:

$$error = \begin{cases} \dfrac{e^2(B-1)}{B(e+Bh)} & \text{if } e \leq \dfrac{B}{1+B} \\ \dfrac{h^2B(B-1)}{e+Bh}, & \text{otherwise} \end{cases} \tag{34}$$

Considering the great complexity of the error function above, an empirical study is conducted in this section to analyze the kappa approximation error.

In the literature, the commonly used interpretation of kappa [33] is shown in Table 3. It demonstrates that the interpretation of kappa is defined for each step of 0.2. Therefore, $|error| \leq 0.1$ are considered as good approximation in this study.

Figure 2 shows 3D plot of function $|error(e, B)|$ for $0.5 \leq B \leq 2$. As can been seen, when $B = 1$, $error = 0$. Moreover, when B is fixed, $|error|$ becomes higher when $e \to \frac{B}{1+B}$; when e is fixed, $|error|$ tends to smaller when $B \to 1$. Empirically, for $|error| \leq 0.1$, we get e and B as follows:

1) $0 \leq e \leq 0.53, B \geq 0.67$;
2) $0.53 \leq e \leq 0.75, 0.5 < B \leq 1.5$;
3) $0.75 \leq e \leq 0.95, 0.5 < B \leq 2$; or
4) $0.95 \leq e \leq 1, 0.5 < B \leq 7.87$.

Note the results above only enumerate some (e, B)'s for $|error| \leq 0.1$, other (e, B)'s hold for $|error| \leq 0.1$ as well.

IV. VALIDATION EXPERIMENTS

The relationship in equation (30) introduces error due to the use of kappa approximation in the derivation. As demonstrated in Section III-B, kappa approximation error is affected by both prevalence and bias problems, therefore in this section we will study the effect of bias and prevalence problems on the relationship approximation between kappa and Youden's J in equation (30), respectively. A synthetic dataset is considered and the performance is evaluated by comparing the true relationship between kappa and Youden's J fitted by a linear regression and the relationship approximation.

A. SYNTHETIC DATASET

To generate dataset with N instances satisfying the annotation generation model in Section II-A, a synthetic data generation process is considered. Firstly, the ground truth (i.e. X_2) is generated as follows: a set of N random values are first generated from uniform distribution in the range [0, 1], and then compared with a threshold T_2 ($0 < T_2 < 1$) for category assignment. The label is assigned as category 0 if the corresponding random value is less than the threshold, and category 1 otherwise. It can be easily shown that such process yields $P(X_2 = 0) = T_2$.

Secondly, preset the proportion of easy instances e, the annotation generation process for X_1 is as follows: labels of the last Ne instances are first copied from X_2 to simulate the easy instances, then the remaining labels are generated by a process similar to the X_2 annotation generation process above with threshold T_1. In the end, it yields $P(l = E) = e$, $P(X_1 = 0|l = H) = T_1$, and $P(X_1 = 0|l = E) = T_2$. More importantly, it can be verified that conditions of equations (2), (3) and (4) are satisfied as well.

Finally, the synthetic dataset is generated as follows: 1) X_2 is generated by the X_2 annotation generation process above with $T_2 = q_0$, 2) with a preset e, X_1 is generated by the X_1 annotation generation process above with $T_1 = p_0$, 3) preset different e's and repeat step 2) for each e to simulate the annotations obtained from different annotators, and 4) repeat steps 1)-3) for different (q_0, p_0)'s. In the experiments, $N = 10^5$ and e is set in the range of [0.05, 0.95] with step 0.05, yielding 19 sets of annotations for each (q_0, p_0). (q_0, p_0) is set during experiments for different purposes. Without loss of generality, $q_0 \leq p_0$ is considered in the experiments since $B(q_0, p_0) = B(1 - q_0, 1 - p_0)$.

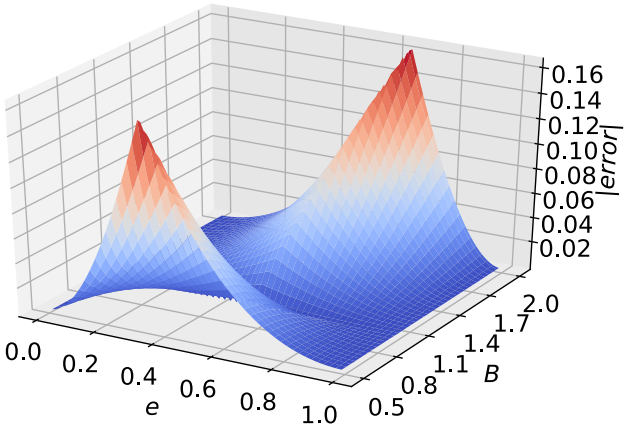

FIGURE 2. 3D plot of function $|error(h, B)|$.

B. PERFORMANCE EVALUATION

To model the relationship between the statistical measures of agreement, a linear regression analysis is conducted. Linear regression models the linear relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). In linear regression analysis, coefficient of determination, denoted by R^2, is used to measure the quality of the fit. The coefficient of determination varies between 0 and 1, where 1 indicates that the model fits the ground truth perfectly.

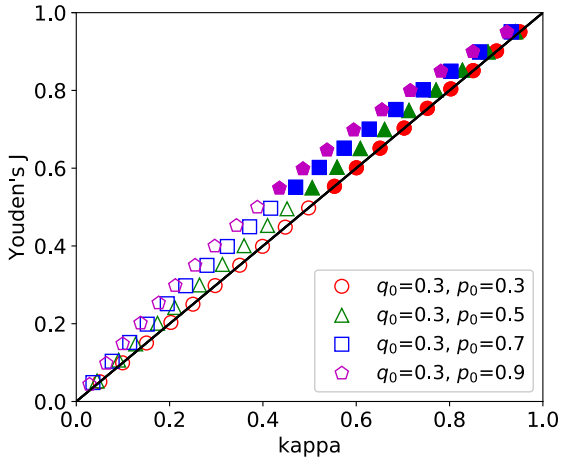

FIGURE 3. Relationship between kappa and Youden's J for eighteen sets of annotations with $q_0 = 0.3$ and $p_0 \in \{0.3, 0.5, 0.7, 0.9\}$.

V. RESULTS

A. THE EFFECT OF BIAS ON THE RELATIONSHIP APPROXIMATION

To study the effect of bias on the relationship approximation, Figure 3 shows the scatter plot of different sets of annotations in the synthetic dataset for $q_0 = 0.3$ and $p_0 \in \{0.3, 0.5, 0.7, 0.9\}$, in which the x- and y-axes are kappa and Youden's J, respectively. In this plot, each point represents the results obtained from a set of annotations, and sets of annotations with same (q_0, p_0) but different e's are denoted by the same shape of markers. Moreover, in this plot, the hollow markers denote the sets of annotations which satisfy condition $e \leq \frac{B}{1+B}$ (denoted as condition #1), and the solid markers represent the others (denoted as condition #2). For better visualization, the line $x = y$ is shown for reference.

As can be seen, for $(q_0, p_0) = (0.3, 0.9)$, the Youden's J increases as kappa increases for the data points marked by hollow pentagons. The same observation can be observed for the data points marked by solid pentagons and other markers.

To quantitatively investigate the effect of bias on the relationship approximation, for data points in Figure 3, linear regression analysis is conducted for each pair of (q_0, p_0) when conditions #1 and #2 are employed, respectively. The results are denoted as "regression" in Table 4 (3rd column), in which coefficient of determination R^2 is shown in the bracket. As can be seen, $R^2 \geq 0.99$ for all linear regression results, indicating that the linear models fit the true relationships perfectly.

Finally, for comparison, the relationship approximation calculated from equation (30) is shown in Table 4 (4th column), which is denoted as "approximation". As can be seen, the relationship approximation is close to the true relationship obtained from linear regression. Moreover, the absolute difference of slope between approximation and regression results for both conditions #1 and #2 become smaller when the bias is less severe; the similar trend can be observed for the absolute difference of intercept between approximation and regression results for condition #1. These results indicate that there is only a negligible approximation error in the simplified relationship when no major bias problems exist.

TABLE 4. Comparison of relationship between kappa and Youden's J for $q_0 = 0.3$ and $p_0 \in \{0.3, 0.5, 0.7, 0.9\}$.

p_0	condition	Regression (R^2)	Approximation
0.3	#1	$y = x$ (1)	$y = x$
	#2	$y = x$ (1)	$y = x$
0.5	#1	$y = 1.12x$ (1)	$y = 1.19x$
	#2	$y = 0.92x + 0.09$ (1)	$y = 0.84x + 0.16$
0.7	#1	$y = 1.22x$ (0.99)	$y = 1.38x$
	#2	$y = 0.85x + 0.17$ (1)	$y = 0.72x + 0.28$
0.9	#1	$y = 1.30x$ (0.99)	$y = 1.57x$
	#2	$y = 0.79x + 0.23$ (1)	$y = 0.64x + 0.36$

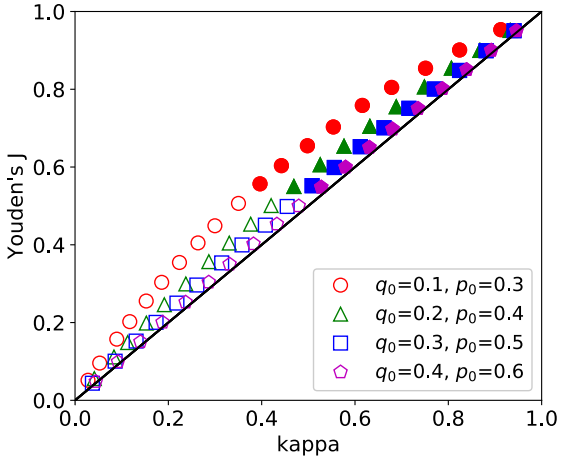

FIGURE 4. Relationship between kappa and Youden's J for eighteen sets of annotations with $q_0 \in \{0.1, 0.2, 0.3, 0.4\}$ and $p_0 = q_0 + 0.2$.

B. THE EFFECT OF PREVALENCE ON THE RELATIONSHIP APPROXIMATION

To study the effect of prevalence on the relationship approximation, Figure 4 shows the scatter plot of different sets of annotations in the synthetic dataset for $q_0 \in \{0.1, 0.2, 0.3, 0.4\}$ and $p_0 = q_0 + 0.2$, in which the x- and y-axes are kappa and Youden's J, respectively. Note to decouple the effect of bias problem, bias is set to be 0.2 in all sets of annotations. Moreover, in this plot, the hollow markers denote the sets of annotations satisfying condition $e \leq \frac{B}{1+B}$ (denoted as condition #1), and the solid markers represent the others (denoted as condition #2). For better visualization, the line $x = y$ is shown for reference.

As can be seen, for $(q_0, p_0) = (0.1, 0.3)$, Youden's J increases as kappa increases for the data points marked by hollow pentagons. The same observation can be observed for the data points marked by solid pentagons and other markers.

The linear regression results for each pair of (q_0, p_0) are shown in Table 5 (3rd column), in which coefficient of determination R^2 is shown in the bracket. For each pair of (q_0, p_0), results are provided for condition #1 and condition #2,

TABLE 5. Comparison of relationship between kappa and Youden's J for $q_0 \in \{0.1, 0.2, 0.3, 0.4\}$ and $p_0 = q_0 + 0.2$.

q_0	condition	Regression (R^2)	Approximation
0.1	#1	$y = 1.43x$ (0.97)	$y = 1.89x$
	#2	$y = 0.69x + 0.33$ (1)	$y = 0.53x + 0.47$
0.2	#1	$y = 1.21x$ (1)	$y = 1.38x$
	#2	$y = 0.85x + 0.17$ (1)	$y = 0.73x + 0.27$
0.3	#1	$y = 1.12x$ (1)	$y = 1.19x$
	#2	$y = 0.92x + 0.09$ (1)	$y = 0.84x + 0.16$
0.4	#1	$y = 1.05x$ (1)	$y = 1.08x$
	#2	$y = 0.96x + 0.04$ (1)	$y = 0.92x + 0.08$

respectively. As can be seen, $R^2 \geq 0.97$ for all of the linear regression results, indicating that the linear models fit the true relationships perfectly.

Finally, for comparison, the relationship approximation calculated from equation (30) is shown in Table 5 (4th column). As can be seen, the relationship approximation is close to the true relationship obtained from linear regression. Moreover, the absolute difference of slope between approximation and regression results for both conditions #1 and #2 become smaller when the prevalence is less severer; the similar results can be observed for the absolute difference of intercept between approximation and regression results for condition #2 as well. These results indicate that there is only a negligible approximation error in the simplified relationship when no major prevalence problems exist.

VI. APPLICATION EXAMPLE: DIABETIC RETINOPATHY DIAGNOSIS

In this section we demonstrate the validity and application of the relationship in equation (30) on a real-life dataset collected from a diabetic retinopathy (DR) screening study [32], as described below.

A. ANNOTATED DR DATASET

The dataset consisted of 1,589 digital fundus images in 45° field-of-view, which were either macula-centered or optic-disk-centered. The images were annotated by a group of seven graders in terms of DR severity according to the International Clinical Diabetic Retinopathy (ICDR) scale: none, mild, moderate, severe, and proliferative DR. On this scale moderate DR and above are clinically categorized as referable DR. Patients diagnosed with referable DR are recommended for further examination. For this study, we consider the binary classification task of detecting referable DR based on the graders' annotation, where moderate, severe and proliferative grades are considered as the positive class and the rest are considered as the negative class. Among the seven graders, three were retinal specialists (G_1, G_2 and G_3) and four were general ophthalmologists (G_4, G_5, G_6, and G_7).

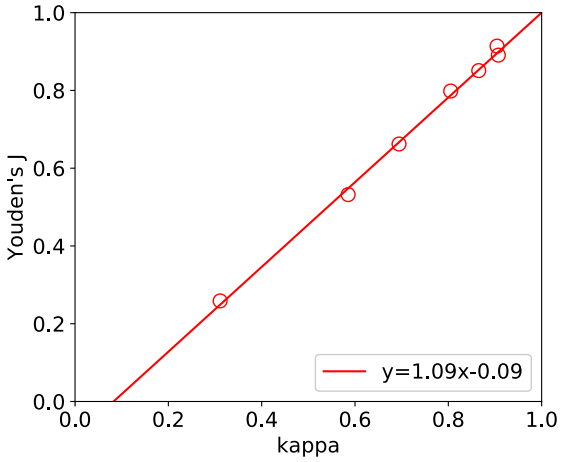

FIGURE 5. Relationship between kappa and Youden's J for seven graders in detecting referable DR.

B. RELATIONSHIP EVALUATION

To verify the validity of the relationship in equation (30), we conducted a linear regression analysis between kappa and Youden's J on the results from the seven graders. For this analysis, a majority voting from G_1, G_2 and G_3 (who were retinal specialists) was used as the ground truth for the DR images. Figure 5 shows a scatter plot of the DR classification results from the seven graders, in which each circle represents the result from a grader. In this plot, the x- and y-axes correspond to kappa and Youden's J, respectively. The linear regression analysis on the scatter points yields $y = 1.09x - 0.09$ with $R^2 = 0.997$. These results indicate that there is an almost perfect linear relationship between kappa and Youden's J from the seven graders. Moreover, since the intercept is nonzero, the relationship for condition #2 applies. In this case, the slope term yields $1/B = 1.09$, and the intercept term yields $(B-1)/B = -0.09$, both of which give the same $B = 100/109$. These results illustrate the agreement of the derived relationship with the real-life dataset.

C. ANNOTATOR QUALITY EVALUATION

Annotator evaluation is an important task in data annotation since poor quality annotation can adversely affect the results [32]. Here we present an application example of the relationship in equation (30) for annotator evaluation when the ground truth is unavailable.

As demonstrated earlier in equation (31), we have $J = \kappa$ in unbiased annotations, and biased annotations will yield a discrepancy between J and κ according to equation (30). Thus, we can use this discrepancy to compare the annotation results from a pair of annotators. Specifically, for annotator $i \in \{1, 2\}$ in the pair, define

$$d_i = |J_i - \kappa|/\kappa \times 100\% \tag{35}$$

where J_i denotes the Youden's J calculated by using annotator i as the ground truth for the annotation results of the

TABLE 6. Annotator evaluation results for pairs of graders.

grader 1	grader 2	κ	J_1	J_2	d_1	d_2
G_1	G_2	0.7782	0.7597	0.8038	2.38%	3.29%
G_1	G_3	0.8167	0.7962	0.8457	2.51%	3.55%
G_1	G_4	0.7660	0.7534	0.7820	1.64%	2.09%
G_1	G_5	0.6509	0.6185	0.7131	5.00%	9.56%
G_1	G_6	0.5529	0.5040	0.7010	8.84%	26.8%
G_1	G_7	0.2863	0.2416	0.6609	15.6%	131.0%

TABLE 7. Kappa coefficients obtained for different pairs of graders.

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
G_1	1	0.7782	0.8167	0.7660	0.6509	0.5529	0.2863
G_2	-	1	0.7991	0.7901	0.7132	0.6080	0.3174
G_3	-	-	1	0.7812	0.7124	0.6193	0.3452
G_4	-	-	-	1	0.7550	0.5903	0.3261
G_5	-	-	-	-	1	0.6460	0.3928
G_6	-	-	-	-	-	1	0.4749
G_7	-	-	-	-	-	-	1

other annotator, and κ is the kappa coefficient between the two graders.

To help understand the discrepancy term in equation (35), let's consider the scenario that the results from annotator 1 are closer to the ground truth than that of annotator 2, i.e., annotator 1 is better than annotator 2. With annotator 1 used as the ground truth, J_1 will be closer to the true performance of annotator 2. As a result, it will yield a smaller d_1 value.

To demonstrate the use of the discrepancy metric in equation (35), we applied it to assess the seven graders in the DR dataset described in Section VI-A above. The results are summarized in Table 6, where G_1 is compared with each of the rest six graders; similar results were also obtained for other grader pairs, but omitted here for brevity.

As can be seen from Table 6, in grader pairs (G_1, G_2), (G_1, G_3), and (G_1, G_4), both d_1 and d_2 are small in value, indicating that both graders in each pair are equally good. On the other hand, in the other three grader pairs, d_1 is notably smaller than d_2, indicating that G_1 is better than the other grader in each pair. In particular, d_2 is much larger than d_1 in (G_1, G_6) and (G_1, G_7), indicating that G_1 is substantially better than G_6 and G_7.

For comparison, we also applied the pairwise kappa method to evaluate the annotators in the DR dataset [32]. In this method, the kappa coefficient is calculated for each possible pair among the seven graders. The results are shown in Table 7. As can be seen, the pairwise kappa values are in the range of [0.7660,0.8167] among G_1, G_2, G_3 and G_4, suggesting that these graders are in a high degree of agreement with each other. On the other hand, the average kappa values are 0.6955, 0.6033, and 0.2285 for G_5, G_6 and G_7, respectively, indicating that G_7 is in the least agreement with the others. Note that these results are consistent with the discrepancy metric results in Table 6.

VII. DISCUSSIONS

Based on the annotation generation model, this study derived an approximation of kappa, and built a simplified, linear relationship between kappa and Youden's J, which is the summation of sensitivity and specificity minus 1. The relationship shows that for fixed B, as kappa increases, Youden's J increases as well. The analysis of B indicates that majority values of the slope in the relationship are in the range of 0.5 and 2, and the large B happens when the severe prevalence and/or bias problems are present. The error analysis for kappa approximation demonstrate that the kappa approximation is less accurate when e is close to 0.5.

The relationship between kappa and Youden's J provides evidence for the kappa interpretation in Table 3. From the relationship in equation (31) for the unbiased annotation, κ is the lower bound of both sensitivity and specificity, and their values are $(\kappa + 1)/2$ if sensitivity and specificity are equal. For example, the almost perfect agreement $\kappa > 0.8$ indicates both sensitivity and specificity are higher than 80% and their values are higher than 90% when sensitivity and specificity are equal. Similarly, the substantial agreement $0.6 < \kappa \leq 0.8$ suggests that both sensitivity and specificity are higher than 60% and their values are higher than 80% but lower than 90% when sensitivity and specificity are equal.

The relationship between kappa and Youden's J was also validated to show a good agreement with annotation results from seven graders in the DR dataset, demonstrating the applicability of the relationship on real-life data. The discrepancy between kappa and Youden's J was demonstrated to be an effective measure for annotator assessment, which yielded consistent results with the traditional pairwise kappa method. Interestingly, it is noted that this discrepancy measure can be applied even when there are only two graders available. In contrast, the traditional pairwise kappa method cannot be applied for only two graders as it requires to compare the kappa values of one grader versus a group of two or more other graders.

VIII. CONCLUSION

Based on an annotation generation model, this study developed a simplified, linear relationship of Cohen's kappa, sensitivity and specificity by employing 1st-order Taylor approximation. The relationship was further simplified by introducing Youden's J statistic, a classification performance summary for binary classification tasks. The analysis on the linear coefficients in the relationship and the approximation error were conducted. A linear regression analysis was applied to evaluate the relationship by using a synthetic dataset. The results show that there is only a negligible approximation error in the simplified relationship when no major bias and prevalence problems exist. The relationship between kappa and Youden's J was also validated to

show a good agreement with real-life dataset collected in DR diagnosis. The discrepancy between kappa and Youden's J was applied for assessment of annotation quality, and was demonstrated to yield consistent results with the traditional pairwise kappa method.

APPENDIXES

APPENDIX A
KAPPA DERIVATION

In this appendix, we demonstrate the derivations of kappa and its range in details. First, the probability of instances in a category c for X_2 is:

$$\begin{aligned} P(X_2 = c) &= P(X_2 = c, l = E) + P(X_2 = c, l = H) \\ &= P(X_2 = c|l = E)P(l = E) \\ &\quad + P(X_2 = c|l = H)P(H = E) \\ &= q_c e + q_c h \\ &= q_c \end{aligned}$$

Similarly, the probability of instances in a category c for X_1 is:

$$\begin{aligned} P(X_1 = c) &= P(X_1 = c, l = E) + P(X_1 = c, l = H) \\ &= P(X_1 = c|l = E)P(l = E) \\ &\quad + P(X_1 = c|l = H)P(H = E) \\ &= q_c e + p_c h \end{aligned}$$

The chance agreement between the two annotators is:

$$\begin{aligned} p_E &= P(X_1 = 0)P(X_2 = 0) + P(X_1 = 1)P(X_2 = 1) \\ &= q_0(q_0\, e + p_0\, h) + q_1(q_1\, e + p_1\, h) \\ &= (1 - 2q_0q_1)e + p_h h \end{aligned}$$

The relative observed agreement between the two annotators is:

$$p_A = p_e e + p_h h = e + p_h h \tag{36}$$

Therefore, Cohen's kappa coefficient is obtained as:

$$\begin{aligned} \kappa &= \frac{p_A - p_E}{1 - p_E} \\ &= \frac{(e + p_h h) - (1 - 2q_0q_1)e - p_h h}{1 - (1 - 2q_0q_1)e - p_h h} \\ &= \frac{2q_0q_1 e}{2q_0q_1 e + (1 - p_h)h} \\ &= \frac{e}{e + \frac{1-p_h}{2q_0q_1}h} \end{aligned}$$

For the range of kappa, if $e = 0$, $\kappa = 0$; otherwise

$$\kappa = \frac{1}{1 + \frac{1-p_h}{2q_0q_1}\frac{h}{e}}$$

Since the two terms in the above kappa expression satisfy

$$\frac{1 - p_h}{2q_0q_1} = \frac{1}{2}\left(\frac{p_0}{q_0} + \frac{p_1}{q_1}\right) > 0.5$$

and

$$\frac{h}{e} \geq 0$$

thus $0 < \kappa \leq 1$. In the end, the range of kappa is

$$0 \leq \kappa \leq 1$$

APPENDIX B
DERIVATION IN SENSITIVITY AND SPECIFICITY

The probability of the two annotators agree on the instances in category c is as follows:

$$\begin{aligned} P(X_1 = c, X_2 = c) &= P(X_1 = c, X_2 = c, l = E) \\ &\quad + P(X_{=}c, X_{=}c, l = H) \\ &= P(X_1 = X_2 = c|l = E)P(l = E) \\ &\quad +, P(X_1 = X_2 = c|l = H)P(l = H) \\ &= P(X_2 = c|l = E)P(l = E) + \\ &\quad + P(X_1 = c|l = H) \\ &\quad \times P(X_2 = c|l = H)P(l = H) \\ &= q_c e + p_c q_c h \end{aligned}$$

APPENDIX C
PROOF OF $|\frac{B-1}{B}e| < 1$

Proof: Since $B = \frac{1-p_h}{2q_0q_1} = \frac{p_0q_0+p_1q_1}{2q_0q_1}$, then

$$\begin{aligned} \frac{B-1}{B} &= \frac{\frac{p_0q_0+p_1q_1}{2q_0q_1} - 1}{\frac{p_0q_0+p_1q_1}{2q_0q_1}} \\ &= \frac{p_0q_0 + p_1q_1 - 2q_0q_1}{p_0q_0 + p_1q_1} \end{aligned}$$

For $0 < q_0 < 1$, we have $p_0q_0 + p_1q_1 > 0$ and $q_0q_1 > 0$. Thus the equation above has:

$$\frac{B-1}{B} < 1 \tag{37}$$

Therefore, we have proved that the upper bound of $\frac{B-1}{B}$ is 1.

The next step is to prove the lower bound of $\frac{B-1}{B}$ is -1. Recall

$$\begin{aligned} \frac{B-1}{B} &= \frac{p_0q_0 + p_1q_1 - 2q_0q_1}{p_0q_0 + p_1q_1} \\ &= \frac{(q_0 - p_0)(2p_0 - 1)}{q_0(2p_0 - 1) - p_0} \end{aligned}$$

If $p_0 = 0.5$, then $\frac{B-1}{B} = 0$, satisfying $\frac{B-1}{B} > -1$. If $p_0 \neq 0.5$, then we have

$$\frac{B-1}{B} = \frac{p_0 - q_0}{p_0 - q_0/(2q_0 - 1)} = \frac{p_0 - q_0}{p_0 - f(q_0)}$$

where $f(q_0) = q_0/(2q_0 - 1)$. $f(q_0)$ has the following properties:

a) if $q_0 > 0.5$, $f(q_0)$ is monotonically decreasing with $f(q_0) > q_0 > 0.5$;

b) if $q_0 < 0.5$, $f(q_0)$ is monotonically increasing with $q_0 > f(q_0) > 0$.

To prove the lower bound, we analyze $\frac{B-1}{B}$ in four different settings with respect to p_0 and q_0 as follows:

1) if $q_0 > 0.5$ and $p_0 > q_0$: since $f(q_0) - p_0$ decreases much faster than $p_0 - q_0$, therefore

$$\frac{B-1}{B} = \frac{p_0 - q_0}{p_0 - f(q_0)} = -\frac{p_0 - q_0}{f(q_0) - p_0} > -1$$

2) if $q_0 > 0.5$ and $p_0 < q_0$: we get $f(q_0) > q_0 > p_0$, therefore,

$$\frac{B-1}{B} = \frac{p_0 - q_0}{p_0 - f(q_0)} = \frac{q_0 - p_0}{f(q_0) - p_0} > 0$$

3) if $q_0 < 0.5$ and $p_0 > q_0$: we get $p_0 > q_0 > f(q_0)$, therefore,

$$\frac{B-1}{B} = \frac{p_0 - q_0}{p_0 - f(q_0)} > 0$$

4) if $q_0 < 0.5$ and $p_0 < q_0$: since $p_0 - f(q_0)$ decreases much faster than $q_0 - p_0$, therefore

$$\frac{B-1}{B} = \frac{p_0 - q_0}{p_0 - f(q_0)} = -\frac{q_0 - p_0}{p_0 - f(q_0)} > -1$$

In conclusion, we have proved $|\frac{B-1}{B}| < 1$. Therefore, $\left|\frac{B-1}{B}e\right| < e < 1$. Done. ■

REFERENCES

[1] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1558–1567, Jul. 2017.
[2] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, and J. Cuadros, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.
[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
[4] J. Wang and Y. Yang, "A context-sensitive deep learning approach for microcalcification detection in mammograms," *Pattern Recognit.* vol. 78, pp. 12–22, Jun. 2018.
[5] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.
[6] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1196–1204.
[7] A. Ghosh, N. Manwani, and P. S. Sastry, "Making risk minimization tolerant to label noise," *Neurocomputing*, vol. 160, pp. 93–107, Jul. 2015.
[8] A. Malossini, E. Blanzieri, and R. T. Ng, "Detecting potential labeling errors in microarrays by data perturbation," *Bioinformatics*, vol. 22, no. 17, pp. 2114–2121, 2006.
[9] B. Frénay, G. de Lannoy, and M. Verleysen, "Label noise-tolerant hidden Markov models for segmentation: Application to ECGs," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Berlin, Germany: Springer, 2011, pp. 455–470.
[10] C. Pelletier, S. Valero, J. Inglada, N. Champion, C. M. Sicre, and G. Dedieu, "Effect of training class label noise on classification performances for land cover mapping with satellite image time series," *Remote Sens.*, vol. 9, no. 2, p. 173, 2017.
[11] D. Dligach, R. Nielsen, and M. Palmer, "To annotate more accurately or to annotate more," in *Proc. 4th Linguistic Annotation Workshop*, 2010, pp. 64–72.
[12] J. Wang, H. Jing, M. N. Wernick, R. M. Nishikawa, and Y. Yang, "Analysis of perceived similarity between pairs of microcalcification clusters in mammograms," *Med. Phys.*, vol. 41, no. 5, 2014, Art. no. 051904.
[13] P. Ruamviboonsuk, K. Teerasuwanajak, M. Tiensuwan, and K. Yuttitham, "Interobserver agreement in the interpretation of single-field digital fundus images for diabetic retinopathy screening," *Ophthalmology*, vol. 113, no. 5, pp. 826–832, 2006.
[14] J. Bootkrajang and A. Kabán, "Classification of mislabelled microarrays using robust sparse logistic regression," *Bioinformatics*, vol. 29, no. 7, pp. 870–877, 2013.
[15] J. Wang, H. Ding, F. A. Bidgoli, B. Zhou, C. Iribarren, S. Molloi, and P. Baldi, "Detecting cardiovascular disease from mammograms with deep learning," *IEEE Trans. Med. Imag.*, vol. 36, no. 5, pp. 1172–1181, May 2017.
[16] R. J. Passonneau and B. Carpenter, "The benefits of a model of annotation," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 311–326, Dec. 2014.
[17] J. Cohen, "A coefficient of agreement for nominal scales," *Edu. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
[18] B. D. Eugenio and M. Glass, "The kappa statistic: A second look," *Comput. Linguistics*, vol. 30, pp. 95–101, Mar. 2004.
[19] F. Velickovski, L. Ceccaroni, R. Marti, F. Burgos, C. Gistau, X. Alsina-Restoy, and J. Roca, "Automated spirometry quality assurance: Supervised learning from multiple experts," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 1, pp. 276–284, Jan. 2018.
[20] T. Byrt, J. Bishop, and J. B. Carlin, "Bias, prevalence and kappa," *J. Clin. Epidemiol.*, vol. 46, no. 5, pp. 423–429, 1993.
[21] J. Carletta, "Assessing agreement on classification tasks: The kappa statistic," *Comput. Linguistics*, vol. 22, no. 2, pp. 249–254, 1996.
[22] W. D. Thompson and S. D. Walter, "A reappraisal of the kappa coefficient," *J. Clin. Epidemiol.*, vol. 41, no. 10, pp. 949–958, Jan. 1988.
[23] M. Feuerman and A. R. Miller, "The kappa statistic as a function of sensitivity and specificity," *Int. J. Math. Edu. Sci. Technol.*, vol. 36, no. 5, pp. 517–527, 2005.
[24] M. Feuerman and A. R. Miller, "Critical points for certain statistical measures of agreement," *Int. J. Math. Edu. Sci. Technol.*, vol. 38, no. 6, pp. 739–748, 2007.
[25] M. Feuerman and A. R. Miller, "Relationships between statistical measures of agreement: Sensitivity, specificity and kappa," *J. Eval. Clin. Pract.*, vol. 14, no. 5, pp. 930–933, 2008.
[26] J. Wang and B. Xia, "Relationships of Cohen's kappa, sensitivity, and specificity for unbiased annotations," in *Proc. 4th Int. Conf. Biomed. Signal Image Process.*, Aug. 2019, pp. 98–101.
[27] M. Aickin, "Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa," *Biometrics*, vol. 46, pp. 293–302, Jun. 1990.
[28] B. B. Klebanov and E. Beigman, "From annotator agreement to noise models," *Comput. Linguistics*, vol. 35, no. 4, pp. 495–503, 2009.
[29] J. Sim and C. C. Wright, "The kappa statistic in reliability studies: Use, interpretation, and sample size requirements," *Phys. Therapy*, vol. 85, no. 3, pp. 257–268, Mar. 2005.
[30] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.
[31] E. F. Schisterman, N. J. Perkins, A. Liu, and H. Bondell, "Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples," *Epidemiology*, vol. 16, pp. 73–81, Jan. 2005.
[32] J. Wang, Y. Bai, and B. Xia, "Feasibility of diagnosing both severity and features of diabetic retinopathy in fundus photography," *IEEE Access*, vol. 7, pp. 102589–102597, 2019.
[33] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: The kappa statistic," *Family Med.*, vol. 37, no. 5, pp. 360–363, 2005.

JUAN WANG received the B.S. and M.S. degrees in electrical engineering from the University of Electronic Science and Technology of China, in 2007 and 2010, respectively, and the Ph.D. degree in electrical engineering from the Illinois Institute of Technology, in 2015. She was an Assistant Specialist with the University of California Irvine, in 2016. She is currently an Information Scientist with Delta Micro Technology Inc., Laguna Hills, CA, USA. Her research interests include computer-aided diagnosis, medical imaging, machine learning, and deep learning.

YONGYI YANG is currently a Harris Perlstein Professor with the Department of Electrical and Computer Engineering, Illinois Institute of Technology. His recent research activities are mostly in computerized techniques for breast cancer detection and diagnosis, and in image reconstruction methods for cardiac diagnostic imaging. His research interests include medical imaging, machine learning, pattern recognition, and biomedical applications. He has authored or coauthored over 250 peer-reviewed publications in these areas. He is a fellow of the American Institute for Medical and Biological Engineering (AIMBE).

BIN XIA received the bachelor's degree from Beijing University, in 1988, and the master's and Ph.D. degrees in physics from the University of Washington, in 1990 and 1996, respectively. From 1997 to 2009, he was a Develop Engineer with Teradyne Inc. From 2010 to 2016, he was the Director of integrated circuit design with Nurotron Biotechnology Inc. He is currently the Co-Founder of Shenzhen SiBionics Company Ltd., and Shenzhen SiBright Company Ltd. His research interests include AI applications in medical instruments.