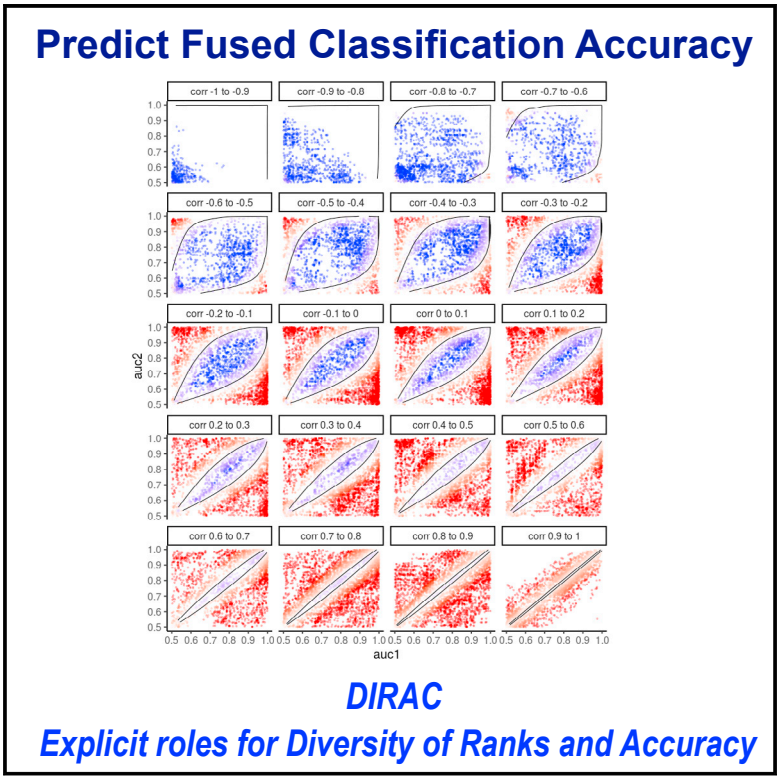
Article

Patterns

Ranks underlie outcome of combining classifiers: Quantitative roles for *DIVERSITY* and *ACCURACY*

Graphical abstract



Highlights

- Several long-standing problems in system fusion for classification problems resolved
- System-level fusion outcome is predictable given characterization of the input models
- The "mechanism" is domain independent and acts at the level of ranks, not scores
- Quantitative demonstration of joint influence of *ACCURACY* and *DIVERSITY* on outcome

Authors

Matthew J. Sniatynski, John A. Shepherd, Thomas Ernst, Lynne R. Wilkens, D. Frank Hsu, Bruce S. Kristal

Correspondence

bkristal@bwh.harvard.edu

In brief

Combining classifier systems potentially improves predictive accuracy, but outcomes have not been predictable. Classification most commonly improves when the classifiers are sufficiently "good" ("*ACCURACY*") and "different" ("*DIVERSITY*"), but the specific influence of these factors on outcome remains unknown. We develop the DIRAC framework (*DIVERSITY* of Ranks and *ACCURACY*), which accurately predicts outcome of both score-based fusions (from exponentially modified Gaussians) and distribution-independent, rank-based fusions. DIRAC was validated using biological imaging data. DIRAC itself is domain independent and has broad expected utility.

Sniatynski et al., 2022, Patterns *3*, 100415
February 11, 2022
https://doi.org/10.1016/j.patter.2021.100415

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Ranks underlie outcome of combining classifiers: Quantitative roles for *DIVERSITY* and *ACCURACY*

Matthew J. Sniatynski,[1,2] John A. Shepherd,[3,7] Thomas Ernst,[4,8] Lynne R. Wilkens,[5] D. Frank Hsu,[6] and Bruce S. Kristal[1,2,9,*]

[1]Division of Sleep and Circadian Disorders, Department of Medicine, Brigham and Women's Hospital, 221 Longwood Avenue, LM322B, Boston, MA 02115, USA
[2]Division of Sleep Medicine, Harvard Medical School, Boston, MA 02115, USA
[3]School of Medicine, University of California San Francisco, San Francisco, CA 94143, USA
[4]John A. Burns School of Medicine, University of Hawaii at Mānoa, Honolulu, HI 96813, USA
[5]University of Hawaii Cancer Center, University of Hawaii at Mānoa, Honolulu, HI 96813, USA
[6]Department of Computer and Information Science, Fordham University, LL813, 113 West 60th Street, New York, NY 10023, USA
[7]Present address: Population Sciences in the Pacific Program (Cancer Epidemiology), University of Hawaii Cancer Center, Honolulu, HI 96813, USA
[8]Present address: Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, USA
[9]Lead contact
*Correspondence: bkristal@bwh.harvard.edu
https://doi.org/10.1016/j.patter.2021.100415

THE BIGGER PICTURE It can be advantageous to combine multiple predictive models for power or robustness, but it is recognized that realizing these potential gains cannot be guaranteed, especially when the input models cannot be appropriately weighted *a priori* or the resulting fusion models cannot be cross-validated. This, and a series of mathematically related problems in different guises, fundamentally limits the ability to optimally use all available models to improve classification across essentially all domains in which more than one potentially useful model exists. We show that any fusion's outcome is fully predictable/explicable given characterization of the models to be fused. The "mechanism" described acts at the level of ranks, not scores, which extends our findings to all distributions and, functionally, to any domain of interest. We are elucidating the underlying math, following the framework's implications for data science, and using this approach on real-world problems.

1 2 3 4 5 **Concept:** Basic principles of a new data science output observed and reported

SUMMARY

Combining classifier systems potentially improves predictive accuracy, but outcomes have proven impossible to predict. Classification most commonly improves when the classifiers are "sufficiently good" (generalized as "*ACCURACY*") and "sufficiently different" (generalized as "*DIVERSITY*"), but the individual and joint quantitative influence of these factors on the final outcome remains unknown. We resolve these issues. Beginning with simulated data, we develop the DIRAC framework (*DIVERSITY* of Ranks and *ACCURACY*), which accurately predicts outcome of both score-based fusions originating from exponentially modified Gaussian distributions and rank-based fusions, which are inherently distribution independent. DIRAC was validated using biological dual-energy X-ray absorption and magnetic resonance imaging data. The DIRAC framework is domain independent and has expected utility in far-ranging areas such as clinical biomarker development/personalized medicine, clinical trial enrollment, insurance pricing, portfolio management, and sensor optimization.

INTRODUCTION

Limitations in our ability to optimally combine the results of prediction and/or classification approaches, without relying on separate/additional validation datasets, represents a fundamental and largely accepted limitation affecting the data-driven analyses and sophisticated modeling approaches that have revolutionized modern science, medicine, and business.

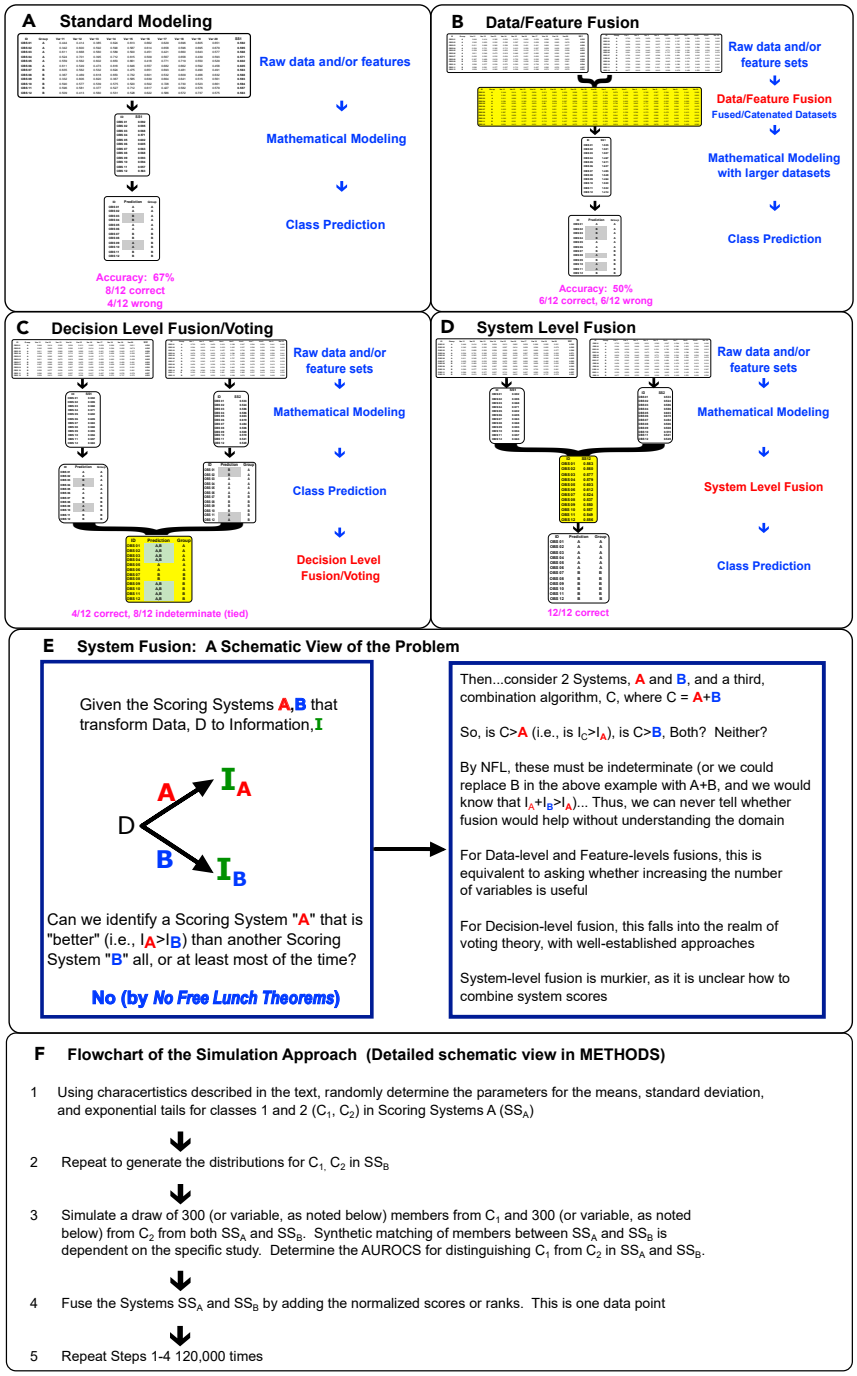


Figure 1. Schematics showing fusion levels and the simulation approach used

(A–D) Schematically presented levels of data and analysis. The top layer(s) represent(s) "raw data and/or features." The following main layer is the output of a scoring system, e.g., the score following a simple statistical model such as regression. We note that the scoring system may be as simple as "report the value of variable n," and matches the observation ID to its cognate score. The third main layer is the class prediction list, i.e., whether, based on the score above, the observation (and cognate ID) is assigned to class 1 (C_1) or class 2 (C_2). The potential levels of fusion (or lack thereof in A) are labeled in red, highlighted in yellow, and denoted by the rotated brace ({) in (B) to (D).

(E) Overview of why system-level fusion has been problematic.

(F) Basic flowchart of the initial simulations used in this study.

Details and additional schematics are provided in experimental procedures. The schema from (A) to (D) parallels that in Jaafar and Ramli,[10] but the nomenclature scheme parallels Ng and Kantor.[14]

The modeling methods applied in these domains span the breadth of statistical and mathematical knowledge, from straightforward parametric methods such as linear and logistic regression, through more complex non-parametric methods such as random forest classification and support vector machines, to ensemble classifiers and deep-learning neural networks. With sufficient data and in well-characterized domains, optimal modeling and analysis approaches drawn from the above are well studied. Where these conditions cannot be met, it is well recognized that determining the optimal modeling approach may be impossible. Such limitations also often make it impossible to build single models that optimally capture the richness of a given dataset. A potential alternative lies in the well-recognized observation that combining multiple different prediction/classification approaches may sometimes improve overall accuracy. Our inability to know when such combinations are useful—and thus to take advantage of the individual systems' strengths and compensate for their weaknesses—is due to an incomplete understanding of how such combination approaches may act to improve overall accuracy.

Combining multiple, weaker models has the potential to improve outcomes. Similar to personal opinions and judgments, different data sources and mathematical models may agree, partially agree, or totally disagree with each other. The potential utility—and complexity—of taking multiple points of view into account has long been recognized, and has been formally considered since at least the 1700s.[1–9] Today, the overall approach itself is known by many names, including information fusion (IF), and the specifics of such combination processes are as varied as they are widespread. These processes fall into one of three general categories (Figures 1A–1D), which we refer to as data (or feature) fusion, system fusion, and decision fusion,[10–13] but all aim to improve output prediction accuracy. We begin by describing the distinctions between these categories, along with their specific advantages and disadvantages.

Data (or feature) fusion is the most straightforward of the three fusion categories and can be thought of as a concatenation of two or more sets of measurements or engineered features pertaining to the same population of samples (e.g., concatenating a population's gene expression levels with a dataset of corresponding clinical phenotypes). Concatenation itself is conceptually well understood,[13] and the resulting concatenated dataset is potentially richer and amenable to all of the familiar statistical/informatics analysis approaches already outlined, such as regression, classification trees, and ensemble classifiers. To succeed, this approach requires the analyst to identify appropriate scaling and appropriate weighting (e.g., millions of genetic markers versus a few clinical variables), and must also address other mismatch problems such as categorical versus continuous variables, all of which potentially complicate model choice and increase model complexity. Data fusion approaches also inherently worsen N versus P problems (number of observations versus number of measured variables), increasing the chance of overfitting, and requiring large sample numbers (often prohibitively large) for adequate training, testing, and validation.

Decision fusion approaches, such as voting, operate at the other end of the spectrum from data fusion and involve the integration of the final decisions of a set of mathematical models. For example, if a pool of different classifier systems is to predict a patient's disease status (positive/negative), decision fusion involves combining the positive/negative votes of each classifier into a final output decision. Decision-level fusion approaches have been studied in depth, particularly in the context of voting theory,[5,15] and remain an active area of research[16] Decision fusion approaches inherently circumvent the scaling, weighting, and mismatch problems inherent to data fusion, but have limited ability to address diverse opinions, can create ties, and cannot naturally account for the certainty/confidence of individual classifiers, although approaches to all of these have been proposed.[4]

System fusion combines mathematical models at a level between data fusion and decision fusion: after the models have been derived from the input data but before a classification decision has been reached. This involves combining appropriately scaled intermediate output of the classification systems in question before this output is transformed into a decision. Differing data types or measurement distributions are accommodated by training separate specialized systems on each, eliminating mismatch problems, and the scaled scores assigned by these systems typically reflect the certainty of the individual sample predictions. As a result, system fusion has the potential to be more flexible and powerful than data or decision fusion, but as a general category it is also the least well understood.

It has been recognized that system fusion performance (i.e., whether fusion improves accuracy or not) is determined, at least in part, by the accuracy of the initial systems and by the diversity between (or among) them.[3,5,6,17–20] For the accuracy of the fused system to be an improvement over that of its constituents, it seems intuitive—and is generally accepted—that these constituents must be both "good" enough to begin with and "different" enough to be worth combining. Despite decades of work, the quantitative details of the relationship between accuracy and diversity remain elusive.[20,21] Multiple approaches for system fusion have been developed and found to work well in some domains but not in others,[14,22,23] and it is recognized that the generalized problem of determining whether two systems should be fused has no solution. This limitation derives from the "no free lunch" (NFL) theorems,[24–28] which were formalized in the 1990s in the context of optimization theory and machine learning. The NFL theorems indirectly indicate that, in the general case (across all possible application domains), the benefit, or lack thereof, derived from the fusion of two models is inherently unknowable; there is no optimal algorithmic approach in the absence of context (Figure 1E). The mathematical implication of this result for our report is that, in the absence of context, it is impossible to tell whether fusing the output of two systems will be beneficial.

Given this general constraint on predictability in the absence of context, it is important to note that some understanding[14,23] has been gained in domain-specific subclasses of information fusion, such as the use of linear combinations and rank-score diversity in information retrieval[29–31] and rank-score diversity in *in silico* drug screening.[32–34] In both these domains the signal of interest tends to lie in a single tail of the distributions of the model outputs, giving a probabilistic structure that can guide fusion approaches and circumvent the NFL limitations. Accuracy improvements in these approaches have been typically in the 70% range, although isolated fully accurate predictions have been observed.[34,35]

In the current study, rather than focusing on a domain-specific context (e.g., information retrieval) to understand and improve system-level fusion, we took the converse approach. Specifically, we sought to identify general-purpose signatures of context to find another route to circumvent NFL's limitations. We thus focused on one particular type of system fusion across as many different input data distributions as possible, so as to identify characteristics of the input systems and their data that may identify in advance situations where fusion is likely to be beneficial. The systems that we focused on were quantile classification systems similar to those that we commonly encounter in our biological work. These assign a monotonically increasing score to each sample to be classified and measure class separation performance using the area under the receiver-operating characteristic curve (AUROC). We measured the diversity between these systems using common correlation metrics (e.g., Pearson correlation) and restricted ourselves to the examination of pairwise fusions only, performing these by averaging the scores of the two systems across each sample.

The results were compelling, revealing a precise, quantitative relationship between ***DIVERSITY*** and ***ACCURACY***, which held in both simulated and real-world biological data (regarding notation, we hereafter use bolded, italicized, small-capitals ***DIVERSITY*** and ***ACCURACY*** to refer to these as generalized mathematical concepts, and normal-font "diversity" and "accuracy" to refer to these in specific instances). Secondary analysis revealed that the relationship observed is solely dependent on the rankings of the samples in the classification systems and is not a direct result of their scores. We refer to this relationship as the DIRAC framework (*DI*VERSITY of Ranks and *AC*CURACY). We present the empirical evidence for the relationships underlying this framework, and discuss its potential implications, applications, and future extensions.

RESULTS

We begin by exhaustively exploring simple averaging of model outputs to determine whether there are conditions in which this

average will consistently outperform the better of the two individual models. This is pairwise system fusion in its most basic instantiation. We use the terms "system" and "scoring system" to represent anything (e.g., analytical test, algorithm, mathematical model) that first gives a single numerical assignment to every sample in a population of samples and then classifies based on that score, for example by thresholding or by separating the top quantile from the bottom quantile. We note that such "systems" can be as complex as fitted ensemble classifiers or as simple as single measurements, such as fasting blood glucose levels. We denote systems as "SS_A," "SS_B," and so forth, and scale their output from 0 to 1. The objective was to study the conditions in which a fused scoring system proved more accurate as a classifier (measured using the AUROC) than either of the progenitor scoring systems.

Simulated data and scoring systems

To determine the relationship between the characteristics of the input systems and the outcome of their fusion without confounding by domain-specific factors, analytical noise, and/or classification errors (such as incorrect labeling), we studied fusion in simulated scoring systems created from probability distributions chosen to approximate those of the data that we typically see in real-world applications (Figures 1F and S18). Commonly, this is a mixture of two Gaussian-family distributions; these distributions, hereafter class 1 (denoted C_1) and class 2 (denoted C_2), may represent, for example, the measurement of a given biomarker in individuals with or without a given disease. Furthermore, standard Gaussian distributions may be combined with exponential distributions to create exponentially modified Gaussian (EMG) distributions, which can have a significantly wider tail, representing many types of real-world measurements with greater fidelity. Each of the simulated scoring systems to be fused thus consists of two Gaussian or EMG distributions that represent the two classes, so each pairwise fusion event features four different distributions: $C_1\_SS_A$, which denotes the distribution corresponding to class 1 within scoring system A, and likewise $C_1\_SS_B$, $C_2\_SS_A$, and $C_2\_SS_B$.

Data samples drawn from the C_1 and C_2 distributions from one scoring system (e.g., $C_1\_SS_A$, $C_2\_SS_A$) can be interpreted as scores for the purpose of classification, with the score of a single data sample within a single scoring system, e.g., SS_A, reflecting the likelihood of that sample having originated from C_1 or C_2. Altering the relative difference between the means of C_1 and C_2 will affect this likelihood, as will altering the standard deviations. When the C_1 and C_2 distributions are fully separated, it represents a perfect classifier (i.e., all the C_1 scores will be less than the C_2 scores and AUROC = 1), and conversely, when there is nearly complete overlap, the classifier will be closer to a random guess and AUROC = 0.5. The greater the probability distribution separation between C_1 and C_2 along this continuum, the higher the accuracy (AUROC) of the corresponding scoring system. The AUROC is a generally accepted and useful metric of classifier performance, as it captures the tradeoff between sensitivity and specificity without requiring the selection of a specific threshold, which is required for other performance metrics such as misclassification rate. AUROC is thus a single number that captures overall classifier performance; given this utility and AUROC's ubiquity, it was the metric that we focused on for measuring scoring system accuracy.

Performance of fusions of simulated scoring systems: A quantitative role for *ACCURACY*

To explore the effects of the C_1/C_2 distribution parameters on pairwise fusion performance, a large pool of simulated scoring systems was created, with the means, standard deviations, and exponential scale parameters sampled randomly from uniform distributions (see parameters in experimental procedures). As noted above, the random variations in the difference between the means of C_1 and C_2, coupled with differences in their standard deviations, led to a broad distribution of AUROCs (i.e., accuracy). Randomly selected pairs of scoring systems (SS_A, SS_B) were then fused by averaging individual points (equivalent to the scores from a synthetic sample) from $C_1\_SS_A$ with those from $C_1\_SS_B$, and from $C_2\_SS_A$ with those from $C_2\_SS_B$, and calculating the mean score of each pair. It is important to note that the random sampling inherent to this simulation procedure produces scoring systems that are relatively uncorrelated, because the magnitude of a simulated sample's SS_A score carries no information about that sample's SS_B score. In a real-world scenario a similar uncorrelated pair of scoring systems might be a blood glucose level and a body weight. We will refer to the AUROCs of SS_A and SS_B as $AUROC_A$ and $AUROC_B$, respectively, and $AUROC_M$ as the AUROC of the superior input classifier (i.e., $\max[AUROC_A, AUROC_B]$). The AUROC of the fused system (hereafter, $AUROC_{SF[AB]}$ for the score fusion of SS_A and SS_B) was measured, and compared with $AUROC_M$, specifically $\Delta AUROC_{SF[AB]} = AUROC_{SF[AB]} - AUROC_M$. This quantity is the change in AUROC of the fused system (either an increase or a decrease) compared with the more accurate member of the unfused pair. Repeating this pairwise fusion process across the large pool of scoring systems allowed us to explore how the C_1/C_2 distribution parameters influence the $AUROC_{SF[AB]}$ of the resulting scoring systems and the $\Delta AUROC_{SF[AB]}$.

A visually striking result emerged when the improvement of the fused classifier systems (i.e., $\Delta AUROC_{SF[AB]} > 0$; presented as binary true/false) was plotted as a function of the two input AUROCs ($AUROC_A$ and $AUROC_B$; synthetic EMG and Gaussian score data fusions [Figure 2]). The space is separated into two distinct regions: one central region where $\Delta AUROC_{SF[AB]} > 0$ (which is notably wider in Gaussian fusions) and a peripheral region where $\Delta AUROC_{SF[AB]} \leq 0$ (Figure 2A). The separation of these two regions is not perfect; between them a relatively narrow band exists where the outcome is uncertain (Figures 2A and 2B). This area of uncertainty is notably wider in the EMG-derived fusions (Figure 2B). Plotting $\Delta AUROC_{SF[AB]}$ directly shows that the area of greatest AUROC improvement lies on a diagonal, symmetric about $AUROC_A = AUROC_B$ (i.e., where the input scoring system performances are equal), and that a relatively wide area of zero-to-vanishingly small improvement separates this area from that where the $AUROC_{SF[AB]}$ is not a net improvement over $AUROC_M$ (Figures 2C and 2D). These data provide qualitative evidence that, in the context of a uniformly high level of diversity (correlation near zero), there exists a strong relationship between $AUROC_A$, $AUROC_B$, and $AUROC_{SF[AB]}$.

Performance of fusions of simulated scoring systems: A quantitative role for *DIVERSITY*

As noted earlier, it is generally accepted that when combining two scoring systems, the resulting accuracy is typically better

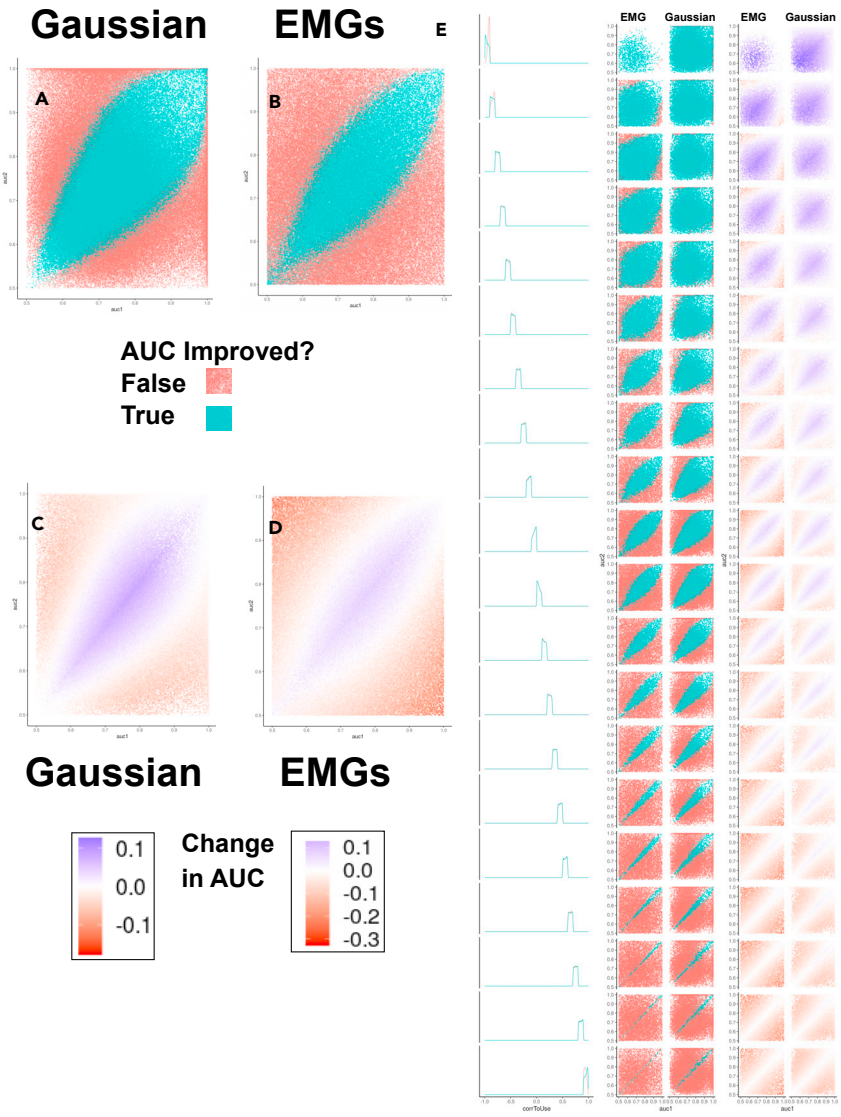


Figure 2. Success or failure of Gaussian and EMG score fusions is predictable

(A–D) Binary (A and B) and continuous (C and D) fusion outcomes of pairwise Gaussian (A and C) and EMG (B and D) score fusions (N = 300/class). Each point represents one fusion and its outcome. The x and y axes indicate $AUROC_A$ and $AUROC_B$. The color indicates the net improvement of the fusion AUROC ($AUROC_{SF[AB]}$), versus $AUROC_M$. In (A) and (B), blue is positive ($\Delta AUROC_{SF[AB]} > 0$) and red is no (or negative) improvement ($\Delta AUROC_{SF[AB]} \leq 0$); (C) and (D) show continuous $\Delta AUROC_{SF[AB]}$.

(E) Each row restricts the within-class Pearson correlation between the fused pair to a 0.1 unit portion of the range from −1 to 1. Moving left to right, the first column shows the distribution of correlations in each slice, the next two columns present binary results for Gaussian and EMG distributions, and the last two columns present the continuous fusion outcomes.

when the input systems are both relatively accurate and also diverse. Having identified a role for ***ACCURACY*** in terms of the AUROCs, we next investigated the role of ***DIVERSITY*** on the results of pairwise scoring system fusions. As a quantitative measure of ***DIVERSITY***, we used the average of the C_1 and C_2 Pearson correlations between any two scoring systems SS_A and SS_B (i.e., avg[Pearson(C_1_SS_A, C_1_SS_B), Pearson(C_2_SS_A, C_2_SS_B)]), and we refer to this quantity as the "within-class" correlation and denote it as D_{PC}(SSA,SSB). The use of within-class correlations was designed to remove the confounding dependence of the overall, or "global" correlation (Pearson(SS_A, SS_B)) on the relative performance of the two scoring systems. For instance, if both SS_A and SS_B have almost perfect AUROC accuracy, it is implicit that SS_A and SS_B must be very similar to each other overall, and therefore have high Pearson correlation. By averaging the correlations specific to each class, this influence of accuracy is removed (there is relevant information in this "global" correlation, which we address in a subsequent report now in preparation; M.J.S. et al., unpublished data).

Using the sampling algorithm outlined above, we altered the within-class pairing of the simulated samples to produce different levels of within-class correlation (see experimental procedures, including schematics). An advantage of using simulated scoring systems like this is that, by explicitly controlling the sample pairing in the generation of the simulated datasets, the correlation between two systems can be varied without changing the mean differences, standard deviations, or any other parameter. This enables an unbiased examination of the effects of correlation alone.

Higher diversity is associated with increased probability that fusion will improve accuracy (i.e., $\Delta AUROC_{SF[AB]} > 0$) and with increases in the maximal improvement resulting from fusion. As the mean within-class SS_A-SS_B correlation increased (Figure 2E, moving top to bottom), the area where $\Delta AUROC_{SF[AB]} > 0$ became increasingly narrow, until at highly positive mean correlation values only a sliver of the originally wide area of improvement remained. The absolute gains in $\Delta AUROC_{SF[AB]}$ at these high correlation values were also reduced, whereas absolute gains were maximized in fusions between strongly negatively correlated scoring systems (Figure 2E, top of rightmost columns).

These results demonstrate that increasing the diversity of the scoring system pair acts everywhere to increase the resulting fusion accuracy ($\Delta AUROC_{SF[AB]}$) and to shift the boundary between positive and negative change outward. This provides quantitative support for the intuitively satisfying notion that, for the fusion of two scoring systems to be worthwhile, they should each contribute at least some different information about the samples under consideration.

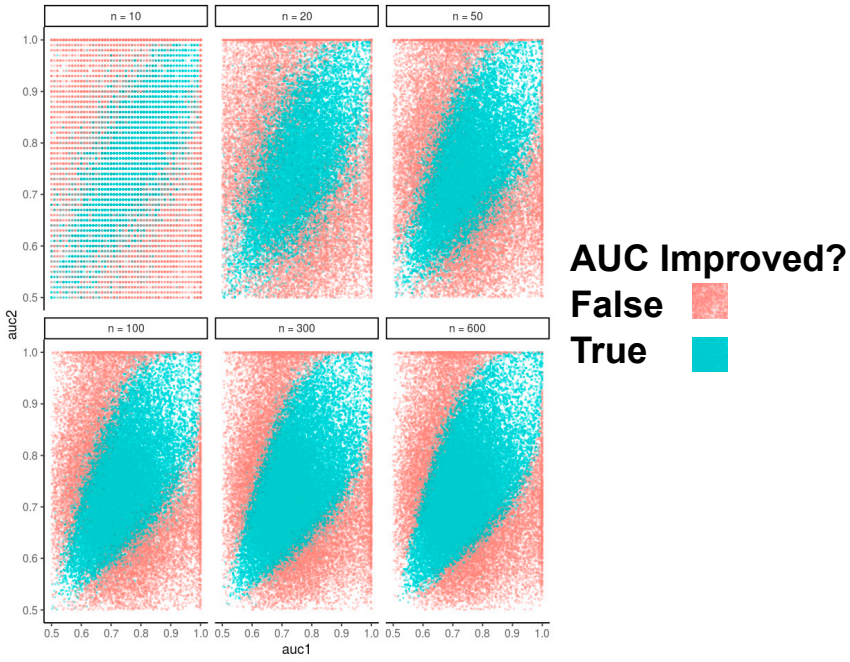


Figure 3. Increasing the number of observations improves prediction precision

Effect of the sample size (N) on the sharpness of the boundary of the region of fusion improvement. The N shown represents the number of samples in C_1 and C_2 (simulated class numbers are balanced). Plotting artifacts are visible in the top left subpanel, where the resolution of the total possible number of AUROC values (with a total balanced N = 20) is smaller than the points representing each fusion in the figure.

Performance of fusions of simulated scoring systems: The effect of N

The number of samples generated for the simulated datasets (the N of the simulation, equivalent to the number of observations in a real-world dataset) affects the accuracy of fusion prediction, as it is directly related to the size of the interval of uncertainty between positive and negative $\Delta AUROC_{SF[AB]}$. We evaluated the effect of sample size N on fusion performance in a series in which each scoring system had 20, 40, 100, 200, 600, or 1,200 total samples split evenly between C_1 and C_2 (Figure 3), and found that increasing N shrinks the region of relative uncertainty between areas of positive and negative $\Delta AUROC_{SF[AB]}$. While it is the accuracy of the two scoring systems and the diversity between them that determines the overall size and shape of the regions of positive and negative $\Delta AUROC_{SF[AB]}$, the data show that the sample size N affects how distinct these positive and negative areas are from each other.

At this point, we have shown using simulated data that the improvement in accuracy of a pairwise system fusion is determined by three parameters: (1) the pairwise diversity, $D_{PC}(SSA,SSB)$; (2) the accuracy of the first scoring system of the pair, $AUROC_A$; and (3) the accuracy of the second scoring system of the pair, $AUROC_B$. We have further shown that our ability to predict this improvement accurately (most critically near $\Delta AUROC_{SF[AB]} = 0$) is determined by a further parameter, the sample size N.

Performance of fusions of simulated scoring systems is driven by ranks

The success of AUROC as a metric of scoring system ***ACCURACY*** suggested to us that the relationship described between ***ACCURACY***, ***DIVERSITY***, dataset size, and fusion performance (change in accuracy and predictability), is mediated by the rankings of the samples and not by their absolute scores. The true-positive and false-positive rates on which the AUROC is based depend only on the ordering of the samples induced by the associated scores. This convinced us to extend our study of pairwise system fusion to focus on analysis of ranking systems. Ranking systems are created by replacing the score of each sample with its corresponding ranking within the dataset under consideration. Two further factors propelled our study in this direction. First, ranking systems have been examined in previous fusion studies, including several that identified conditions in which ranking system fusions outperformed scoring system fusions.[34,36–38] Second, rankings themselves are of particular utility in situations such as ranking candidates (e.g., by risk) for clinical trial enrollment, ranking a list of potential stocks to include in a portfolio, or using rank-based statistics such as the Mann-Whitney U (MWU) test, in situations involving data that are not normally distributed. Previous work has highlighted the equivalence of the MWU test and the AUROC, meaning that pairwise fusions that improve the AUROC inherently improve statistical significance and power.[39]

We repeated the analyses as described above (Figure 2) but using the ranking systems corresponding to each scoring system. This yielded nearly identical plots (Figure 4), with the notable difference that the boundary between the areas of positive and negative $\Delta AUROC_{RF[AB]}$ (the rank equivalent of $\Delta AUROC_{SF[AB]}$) is much sharper, considering an equivalent sample size (N) and within-class correlation range. This results in a generally better ability to predict the accuracy change of a pairwise fusion; visual comparison of scoring and ranking system fusions suggests the improved discrimination at the boundary approximates a 10-fold increase in the sample size. We note that the rank-based equivalent of Pearson correlation is the Spearman rank correlation, which is scaled similarly (1 for perfect correlation to −1 for perfect anticorrelation). Pearson correlation on discrete integer ranks yields the Spearman rank correlation directly (i.e., $D_{PC}(RS_A,RS_B) = D_{SR}(RS_A,RS_B)$).

These results provide evidence that the geometric relationship observed in our initial studies of scoring system fusion and, by extension, the relationship that we propose between ***ACCURACY*** and ***DIVERSITY***, is a manifestation of structure at the level of the sample rankings only (i.e., the ordering imposed by SS_A and SS_B on [the observations in] C_1 and C_2.), and do not involve the absolute or relative values of the sample scores directly.

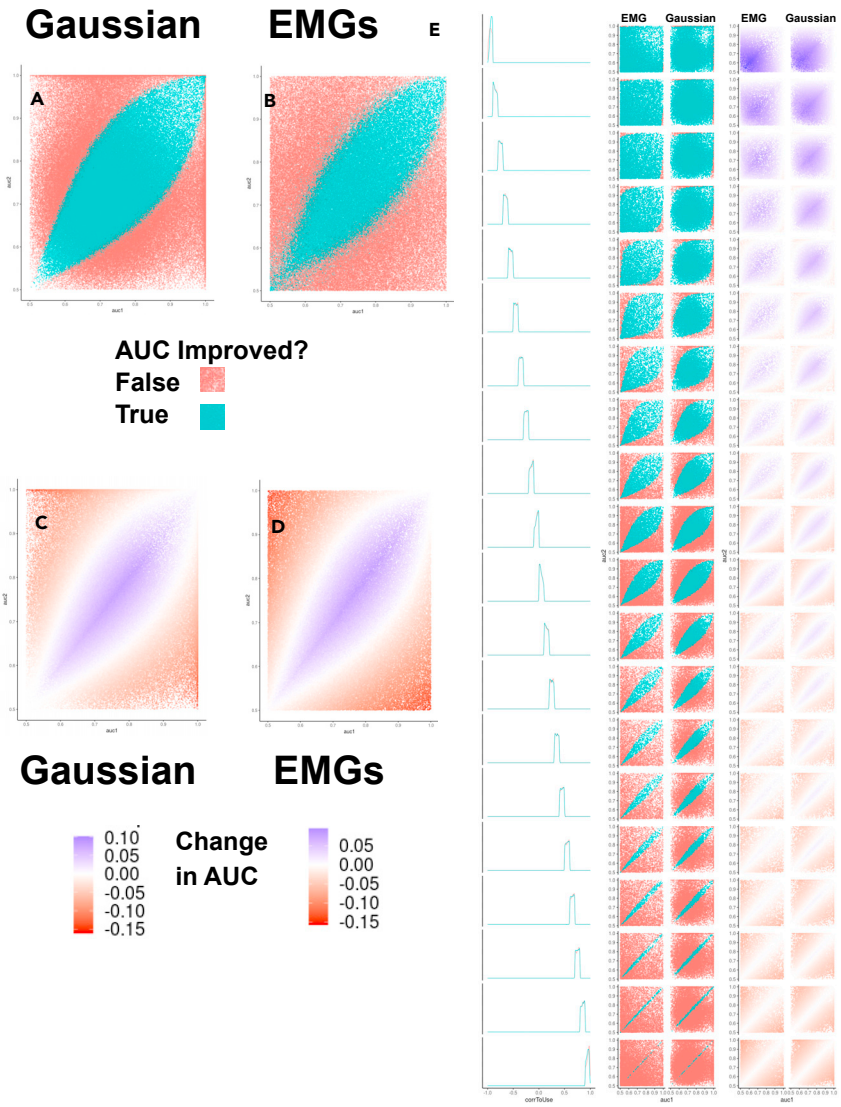


Figure 4. Success or failure of rank fusions is more predictable than score fusions

The analysis shown in Figure 2 is repeated using the rank data.

(A–D) Binary (A and B) and continuous (C and D) fusion outcomes of pairwise Gaussian (A and C) and EMG (B and D) rank fusions (N = 300/class). Each point represents one fusion and its outcome. The x and y axes indicate $AUROC_A$ and $AUROC_B$. The color indicates the net improvement of the fusion AUROC ($AUROC_{SF[AB]}$) versus $AUROC_M$. In (A) and (B), blue is positive ($\Delta AUROC_{SF[AB]} > 0$) and red is no (or negative) improvement ($\Delta AUROC_{SF[AB]} \leq 0$); (C) and (D) show continuous $\Delta AUROC_{SF[AB]}$.

(E) Each row restricts the within-class Pearson correlation between the fused pair to a 0.1 unit portion of the range from −1 to 1. Moving left to right, the first column shows the distribution of correlations in each slice, the next two columns present binary results for Gaussian and EMG distributions, and the last two columns present the continuous fusion outcomes.

LOWESS curves capture the boundary between beneficial and non-beneficial fusions

To validate our simulated data findings and to enable the testing of our ability to predict fusions in the real world, we sought a compact representation of the boundary separating positive and negative $\Delta AUROC_{RF[AB]}$. Describing this boundary is complicated by (1) the variability of the shape of the boundary and (2) the lack of a concise equation describing this boundary (in terms of input accuracy [AUROC] and diversity [Pearson/Spearman]), or a full mathematical description of the pairwise fusion characteristics that give rise to it. We therefore elected to leverage the ability of our simulation framework to generate large amounts of very-high-resolution (high N) data and model the boundary non-parametrically using LOWESS curves. LOWESS (locally weighted scatterplot smoothing also known as LOESS [locally weighted smoothing]) curves model relationships between variables without assuming any mathematical structure and, when trained on our simulated data, produced an excellent approximation of the boundary, correctly separating the vast majority of positive and negative $\Delta AUROC_{SF[AB]}$ and $\Delta AUROC_{RF[AB]}$ (Figure 5).

The DIRAC framework predicts fusions of real-world data

We then examined pairwise system fusion in a real-world situation by applying the DIRAC framework to medical imaging data using the LOWESS curves described above. Medical imaging data were drawn from the Multiethnic Cohort Adiposity Phenotype Study (MEC-APS),[40] a study of adiposity phenotypes in men and women from five ethnic groups, specifically the initial 1,000 subjects recruited (533 women) consisting of approximately equal numbers of Japanese-Americans, African-Americans, Latino(a)s, native Hawaiians, and Caucasians. In MEC-APS, body distributions were determined using both dual-energy X-ray absorption (DXA) and magnetic resonance imaging (MRI). DXA imaging is cheaper than MRI and more clinically available but is less accurate for determining body fat distribution in the viscera, for which MRI is the gold standard. We tested DIRAC's real-world applicability by systematically assessing DIRAC's ability to "predict" whether or not the fusion of two DXA metrics improved predictions of a given MRI metric versus the better DXA predictor. Independent of specific biological outcome(s), we tested the mathematical question of whether our simulations reflect real-world performance. We first converted 31 different DXA measurements into ranking systems (assigning ranks based on the magnitude of the measurements), and for each of these we calculated AUROC against 39 different binary MRI measurement targets (for lists of variables see Data S1 [DXA] and Data S2 [MRI]). We then carried out pairwise within-class Spearman correlation measurement, followed by

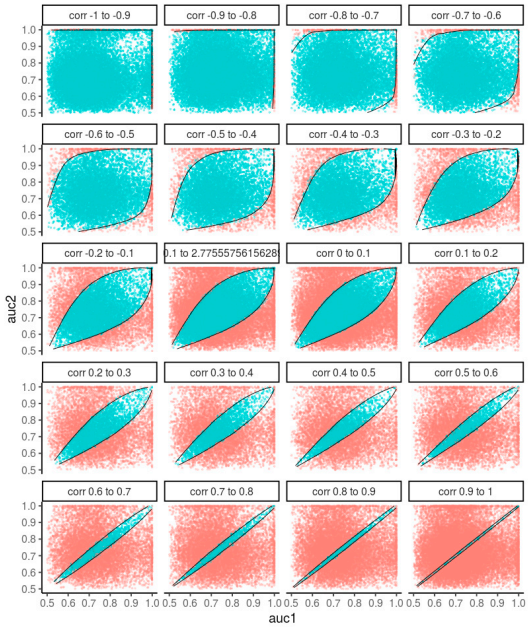


Figure 5. LOWESS curves accurately capture the boundary in rank fusions of simulated data

LOWESS (locally weighted scatterplot smoothing) curves were fit to the rank fusion training data, stratified by correlation into 20 bins as before ($r = -1$ to 1, upper left to lower right, with 0.1 unit intervals).

system fusion, on all possible pairings of DXA measurements, in a manner identical to our simulated data experiments, calculating the $\Delta AUROC_{SF[AB]}$, $\Delta AUROC_{RF[AB]}$, $AUROC_{SF[AB]}$, and $AUROC_{RF[AB]}$ of the fused predictor against these same 39 MRI measurement targets (108,810 fusions [54,405 unique]). Figure 6 shows the real-world pairwise fusions colored by positive/negative $\Delta AUROC_{SF[AB]}$, and overplots the LOWESS curves generated from the positive/negative boundary in the simulated data. It is visually evident that the ***ACCURACY/DIVERSITY*** relationship is identical in both simulated and real-world datasets. The boundary uncertainty in both simulated and real data likely derives at least in part from the range of pairwise within-class correlations included in each plot (i.e., the $\Delta R = 0.1$ correlation range); decreasing ΔR would be expected to reduce such uncertainty.

With respect to the specific analyses shown, this MEC-APS analysis demonstrates how fusing predictors, in this case DXA imaging measurements, can improve classification accuracy, yielding more accurate estimates of given ground-truth MRI-based classification groupings (i.e., quantiles). These analyses also highlighted potentially unexpected pairwise relationships for follow-up. In the demonstration above, for example, the DIRAC framework correctly predicted a substantial gain in accuracy predicting extreme quintiles of visceral fat when some metrics of total fat, percent fat, or fat distribution were fused with a bone mineral density measurement (overall, spine, or pelvis; see subchondral bone mineral density/bone mineral content metrics [Figure S1], RANK fusions [Figures S2–S7], and specific SCORE fusions [Figures S8–S14]). This was unexpected (by us) biologically, and led us to identify existing literature linking bone mass and visceral fat.[41,42]

More importantly, this real-world validation of DIRAC has important implications for the generalized understanding of how information fusion for classification works. Under the specific conditions tested, the DIRAC framework can recapture and use >99% of system-level information without referring to the primary data or scores information. This reveals the following: (1) the basic geometric relationship between $AUROC_A$, $AUROC_B$, and the $D_{SR}(SSA,SSB)$ of the ranks of the observations in the dataset holds between simulated data and a large real-world dataset; (2) the specifics of this relationship are essentially perfectly captured by LOWESS curves based solely on simulated rank data, consistent with the argument above that it is ranks that drive the success of the DIRAC framework; and (3) DIRAC quantitatively captures the interaction between ***ACCURACY*** and ***DIVERSITY*** in information fusion for classification.

DISCUSSION

System-level information fusion: An overview

It is possible to view system-level information fusion from several different perspectives, and it is important to recognize that these partially overlapping perspectives often have notably different approaches, goals, assumptions, and nomenclature. A primary nomenclature difference is an ambiguity in the names used for approaches that we identify as system fusion. These terms include (but are not limited to) system fusion, model fusion,[43] information fusion, measurement or confidence-level fusion,[44] and score-level fusion,[44,45] multiple classifier systems (MCS),[45,46] and combinatorial fusion analysis (CFA).[12] A second nomenclature ambiguity involves descriptive terms being used differently by different authors at different times. For example, the term ''information fusion'' is variously used to cover the entire area, but also to refer to many different and specific processes of information combination nested within it that occur at various levels of fusion, including what we call system fusion. For example, consider two studies that define information fusion as occurring at two possible stages, either ''early'' (corresponding to what we call data/feature fusion), or ''late'' (which we term decision fusion). They independently consider the advantages of an ''intermediate'' fusion step between these two points, which is analogous to what we call system fusion in this study. Both are working using neural networks; Arevalo et al.[47] propose using gated multimodal units to fuse intermediate representations learned by the network, and Kim et al.[48] use separate deep-learning networks to construct the intermediate representations and then combine the outputs of these. A third nomenclature difference, which is more of a functional distinction, pertains to whether additional data-driven training or model fitting is involved in the

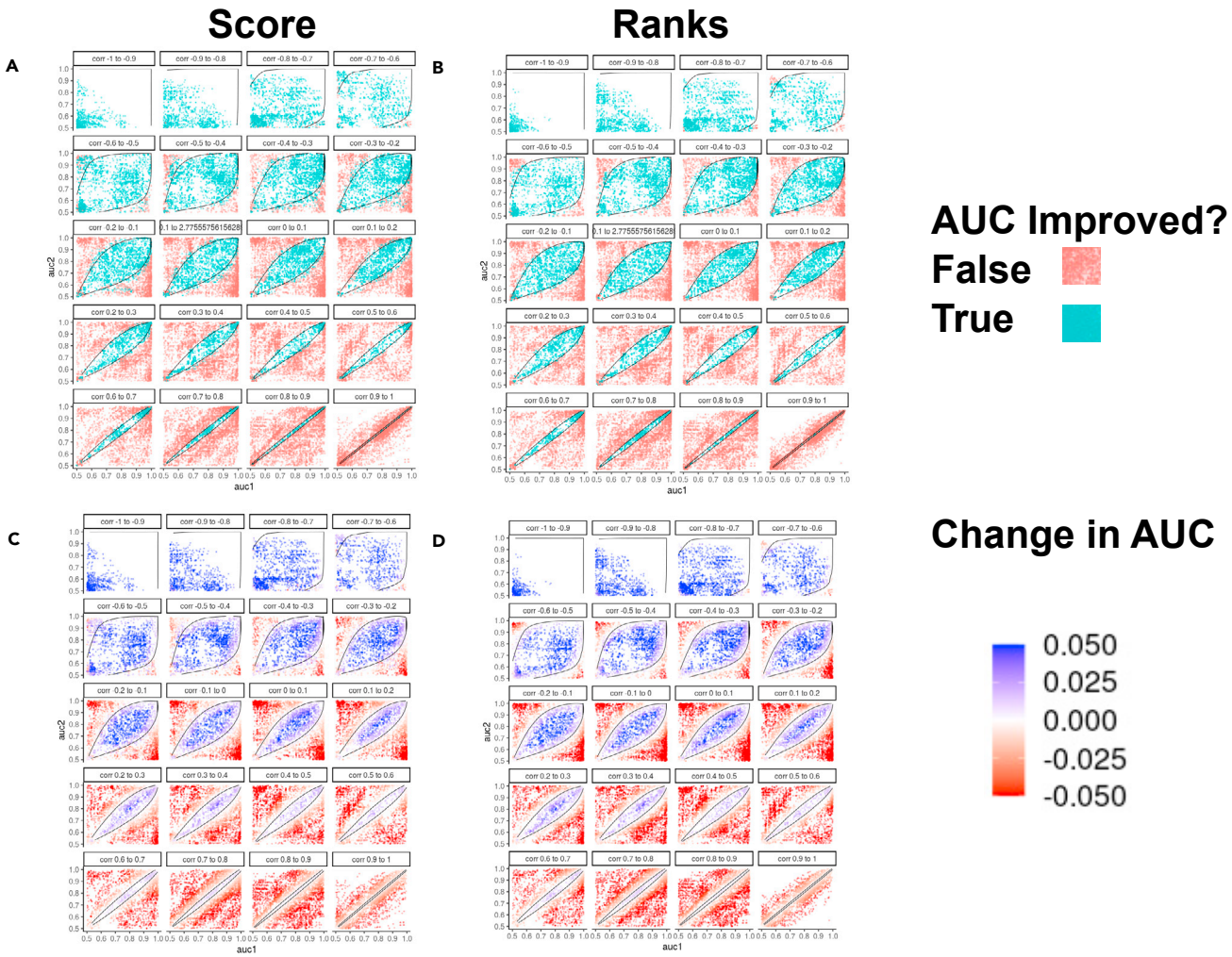


Figure 6. LOWESS curves built on simulated rank fusion data accurately capture the boundary in real-world score and rank fusions

Fusions of two DXA measures to predict the extreme quantiles (tertiles, quartiles, quintiles) of MRI-based measures. Plots are presented in the same way as simulated data in Figures 2, 4, and 5 with the LOWESS curves from Figure 5 superimposed.

(A–D) Binary (A and B, blue is positive [$\Delta AUROC_{SF[AB]} > 0$] and red is no (or negative) improvement [$\Delta AUROC_{SF[AB]} \leq 0$]) and continuous (C and D) fusion outcomes of score (A and C) and rank-based (B and D) fusions. Continuous fusion panels are winsorized at ±0.05 for visual clarity. Equivalent panels without winsorization and winsorized at ±0.2 and ±0.1 are provided as Figures S15, S16, and S17 (respectively, these are versions of this figure at full scale, at −0.2 to 0.2, and at −0.1 to 0.1 for comparison). Each point represents one fusion and its outcome (i.e., two DXA variables and a single MRI target). For visual clarity (symmetry), each fusion is plotted twice (i.e., once with each system as SS_A and again as SS_B). Data are drawn from 1,000 individuals in the MEC-APS; see text and Lim et al.[40]

system fusion. For instance, in many applications the system fusion acts similarly to the final combination step in an ensemble classifier, where each separate system's outputs are used as inputs to a separate modeling step, trained on additional data to optimize their combination. Several authors use the term ''information fusion'' to refer to what we would consider standard statistical or machine-learning modeling steps, using Bayesian model averaging,[49] elastic-net regularization,[50] or Kalman filters.[51,52] This differing use of vocabulary is by no means incorrect, but underscores the difficulty encountered when summarizing a research area that is only nominally united and exceedingly broad. In contrast, in this study we specifically identify system fusion as the combination of model outputs involving no additional/subsequent data-driven training.

With these semantic distinctions in mind, we will summarize some of the recent history of system-level information fusion research. While this will cover many specific applications and identify several heavily studied areas, it will also reveal that there is currently little in terms of high-level, general, domain-independent analysis—this is the main contribution we are hoping to make with this study. We will summarize this recent research in the context of three overlapping categories: (1) approaches that emphasize fusion as a technique in leveraging application/domain-specific knowledge; (2) approaches that emphasize the manipulation of modeling approaches to achieve or improve system fusion; and (3) approaches that emphasize the logistics of the fusion process itself. Reading across these divisions presents a cross-section of the many areas in which the many

variants of system fusion are playing important roles in addressing information-rich problems.

Examining the history of system-level information fusion from the perspective of application/domain specificity, it becomes apparent that many methods seek to fuse information to take advantage of the users' domain-dependent knowledge of specific complementarity existing between/among certain domain-specific signals. This occurs, for example, in information retrieval[31] in *in silico* drug screening[32,34] and biometric and multibiometric system fusion,[45,46] as well as opinion mining (also known as sentiment analysis)[53] and stock market prediction.[43,46] In information retrieval the goal is to identify a small number of documents that are most relevant, considering a set of query criteria. A system-level fusion that emphasizes a domain-knowledge approach is a natural fit here, as different models can be employed that each specialize in different aspects of the desired output, with system fusion combining them into a single information retrieval system. For example, to identify documents pertaining to the business aspects of professional sports, a model trained to filter business-related documents could be fused with a model trained to filter sports-related documents. With the system fusion appropriately configured, only documents scored highly by both input systems will be flagged as relevant. The information retrieval task has historically motivated a great deal of work in fusion approaches, and this tendency continues into the present day, with recent work such as that from Benham and Culpepper, who examine the risk-reward tradeoffs inherent in the domain, showing that carefully selected rank- or score-fusion approaches may help in discriminating irrelevant from relevant results.[54] The overall situation is similar in *in silico* screening, although it is notable that both the target and the inputs are likely to be extremely specific and quantitative as compared with the broader, more qualitative inputs found in information retrieval. A system-level fusion approach that emphasizes domain knowledge is similarly a natural fit here, as one can fuse different models that each capture different sets of local or global chemical properties.[32,34] In the field of biometrics, the goal is to determine whether the subject is the person of interest or to find the optimal matching of the subject to biometric profiles in a database. Biometrics lends itself to domain-driven fusion at multiple levels; the output of redundant sensors for a given property may be fused (inviting a data fusion approach), and the output of multiple sensors for unrelated properties (e.g., retinal and fingerprint scans) may also be combined, at either a system fusion or a decision fusion level.[10,44,45] Other recent fusion applications emphasizing domain-centric knowledge and local context include target tracking applications[55–57] whereby the local conditions (the weather conditions, for instance) may direct the combination of relevant target information, and medical/diagnostic applications,[58] which take advantage of the fundamentally different types of information available from wearable sensors and from electronic medical records. Other areas where information fusion may be able to take advantage of specific domain-centric knowledge include food quality authentication,[59] construction engineering and management,[60] kiwifruit detection,[61] manufacturing service resource allocation,[62] and wind turbine fault diagnosis,[63] although some of these applications, particularly those deriving from fuzzy set theory, blur the line between the two classes we are referring to as "system fusion" and "decision fusion."

Examining the history of system-level information fusion from the perspective of the creation and manipulation of the input models themselves reveals a set of methods that take a different approach. These methods focus on finding or creating inputs that are most likely to capture distinct (and possibly complementary) signals due to their mathematical nature. In other words, these approaches seek to select and configure models such that the diversity between them is maximized.[46,64] Examples include fusing a projection-based method that extracts linearly independent information from multiple weak signals with a least absolute shrinkage and selection operator (LASSO)-based approach that emphasizes only the strongest variables, or with a random forest model capable of representing non-linear signals. Another proposed alternative is the deliberate use of unstable classifiers.[46] Though occurring at the system fusion level, many of these approaches incorporate a secondary data-driven training step which learns how to optimally combine the inputs (similar to an ensemble classifier) or incorporate the fusion directly into the primary statistical or machine-learning modeling step. Of the recent work that falls into this category, one uses a multiple kernel support vector machine (coupled with Hilbert-Schmidt independence criterion) to combine the results of different algorithms,[65] one uses a neural network,[66] and another uses Gaussian process regression.[67] Another is a stock market prediction study of bagging and boosting approaches for the maximization of diversity in a base pool of classifiers which may then be fused linearly, non-linearly, statistically, or by using separate machine-learning approaches.[46] A different approach attributed to Gao et al. also focuses on diversity of information, in this case using canonical correlation analysis to effectively construct representative low-dimensional features.[68]

Examining the history of system-level information fusion from the perspective of the fusion process itself reveals approaches that seek to amplify useful signals from the input models in a fully domain- and model-agnostic manner by defining goals or optimality criteria that are independent of both the domain of origin and the mathematical nature of the systems being fused. There are at least two broad classes: those that are purely procedural (i.e., run method A, run method B, then perform a specific action to fuse them) and those that attempt to define rules/criteria that determine when and how to fuse systems, based on some domain-independent aspect of the information at hand. In the first class, commonly used fusion approaches include SUM, PRODUCT, and MIN/MAX (e.g., Fierrez et al.,[45,46] Barak et al.[45,46]). Broadly speaking, "SUM"- and "PRODUCT"-based approaches generally seek to leverage concurrence between the input systems (operating similarly to a Boolean "AND"), whereas "MIN/MAX"-type fusions attempt to incorporate each system's specific knowledge into the fused system (operating similarly to a Boolean "OR"). Other approaches attempt to combine the systems differentially, for example by employing various fixed weighting criteria. In the second class of approaches in this category, various rules are defined prescribing when and how system fusion should be carried out. These approaches differ from the first class in this category in that they are "adaptive" and prescribe different fusion approaches for different input information/signals. However, unlike the other fusion approaches discussed previous to this section, these approaches use only domain-independent aspects of the input

information and remain agnostic to the specific models used for processing it. An example of system fusion approaches in this latter class includes those based on rank-score diversity, where the information used to direct system fusion is the relationship between a sample's rank (within a dataset) and its score (assigned by a modeling system).[12,30,32,34] Research within both classes in this category generally emphasizes the study of the fusion procedure itself, and pays comparatively less attention to context/domain or to specifics of upstream modeling/representation. Some foundational work in this area includes a decomposition of ensemble mutual information into "accuracy" and "diversity" components,[69] a geometric treatment of information fusion,[70] and a revisitation of this concept some years later in the context of information retrieval.[71] Recent work includes a recent survey of fusion techniques in the medical field,[72] a clinical and sensor fusion approach examining multiple different levels of fusion to create an optimal blood-pressure model,[73] and a review of different fusion methods (also applicable at multiple levels) for use in combining image and textual data.[74] Other recent applications of interest in this area include an adaptive system fusion approach for optimally combining biometric data (and detecting identity fraud) in a dynamic environment,[75] an examination of methods for creating a pool of classifiers that are more likely to fuse with each other beneficially,[76] and a dynamic classifier fusion system whereby the question is both which input systems to fuse and how to do it (including on a sample-by-sample basis).[64]

While these approaches have moved the field forward and have been demonstrably helpful for improving modeling and addressing real-world problems, none of the existing approaches we have surveyed have successfully attained two major goals of information fusion that are alluded to throughout the perspectives discussed above, specifically: (1) providing a quantitative understanding of the roles played by ***DIVERSITY*** and ***ACCURACY***, and the relationship between them, so as to enable development of primary modeling approaches more amenable to downstream system-level fusion; and (2) providing a robust, domain-independent framework for system-level fusion such that relevant signal present in primary models is captured if possible and, in the worst case, not destroyed in the system fusion step. Similarly, two other important issues have also largely remained unaddressed. One issue is that most of the existing approaches are heavily focused on improving prediction of the top hits (e.g., information retrieval, *in silico* screening, biometrics), whereas there has been little if any investigation on the effects of system-level fusion's ability to improve accuracy across the entire distribution. The other issue is the one that arguably encompasses them all—there is, as yet, no truly general, unifying theory describing how and why system fusion approaches work. In this report we have largely addressed the first three, and these set the stage for addressing the fourth (now in preparation; M.J.S. et al., unpublished data).

Fusion, DIRAC, and ranks

The main objective in fusion is to achieve a level of accuracy that is higher than is achieved by either of the input systems alone. The current study shows that, for a particular type of pairwise system fusion, the accuracy of the fused system can be modeled with high precision, knowing only the accuracy of the two input systems and the within-class correlation between them. We have explored the dependence of this result on the size of the dataset and have shown that, although systems may be combined at the score level or at the rank level, rank-level fusion predictions are more accurate. This difference is especially noticeable at smaller sample sizes.

We conjecture that the relationships described by the DIRAC framework between input system accuracy (AUROC), pairwise system diversity (within-class correlation), and fused system accuracy (AUROC) originate directly from the sample rankings and not from their absolute scores. The difference in prediction precision between score and rank fusions (i.e., the different amounts of noise at the boundary) may be because the set of rankings of any given length is of finite size, whereas the set of possible score vectors that can give rise to each of these rankings is infinite. We hypothesize that this one-to-infinity mapping produces the additional noise at the $\Delta AUROC_{SF[AB]} = 0$ boundary, as scoring systems will then fuse with perfect predictability only if their distributions sufficiently resemble the uniform distributions of their corresponding ranking systems (i.e., are well calibrated), which is a rare event within the Gaussian/EMG-based simulation framework. The additional information present in the score distributions is not properly accommodated in the system fusion framework that we describe (and which is implicit in the AUROC metric), and manifests instead in the additional noise/uncertainty observed at the scoring system fusion decision boundary. Given this, it is notable that the data shown in Figures 2, 4, and 6 suggest that rank-based fusions are a very close approximation for score-based fusions under the simulation conditions used and in the real-world data tested. In some exceptions, most noticeably those visible here as better-than-expected fusions in the score plots in the $-0.6 < r < -0.5$ panel, additional information present/embedded/encoded in the scores (versus ranks) engenders unexpectedly improved fusions. Further examination determined that many of these fusions involve a high-accuracy system with a bimodal distribution or very wide central peak; 112/135 (83%) of these systems were positively biased toward females (primarily percent fat). Conversely, the lower-accuracy systems here are positively biased toward males (108/135, 80%, primarily lean mass metrics), suggesting the possibility that sex drives the within-class anticorrelation and the fusion essentially attenuates/removes sex as a source of systemic noise in the classification.

The distributions of the scores, or functions involving both scores and rankings, may encode domain-specific information that may be exploited for additional performance in certain situations, although the NFL theorems preclude this from being true generally. This is a promising area for future study. For now, the important point to note is that if fusion prediction relies only on the rankings of the input systems, there is no longer any connection to the original score distributions. This means that the results we present are, by definition, general and domain independent; the only restriction is that the input systems be monotonically increasing scoring functions.

Relative importance of *ACCURACY* versus *DIVERSITY*

Whether considering scoring systems or ranking systems, the distinct regions of positive and negative $\Delta AUROC_{SF[AB]}$ visible in the figures herein make explicit the relationship between

ACCURACY and ***DIVERSITY*** in this general context. ***ACCURACY*** and ***DIVERSITY*** have long been hypothesized to be of fundamental importance to fusion in general, but the relationship has been poorly understood. Our analysis of pairwise fusions in simulated data and our validation of this framework with real-world data has provided quantitative support for the intuition that improvements in accuracy occur more readily when fusing two uncorrelated systems (i.e., with an average within-class correlation near zero). It has further revealed the more surprising reality that fusing negatively correlated system pairs is even more likely to produce increased accuracy, and that the increase itself is likely to be larger, boosting even marginal individual systems to higher (and occasionally very good) performance. Provided both systems in a pairwise fusion have some predictive value, a strongly negative correlation can allow an already accurate system to fuse beneficially with a poor one. In the limiting case, a nearly perfect negative correlation can fuse to perfection (AUROC = 1.0) two systems that are only minimally better than random chance.

Data in the figures make explicit that it can be more beneficial to choose two less accurate systems to fuse than two more accurate systems if and only if the correlation is lower (ideally negative) for the former pair than for the latter. Conceptually, systems with high within-class correlation are typically understood to be representative of the same latent signal in the data with respect to the target outcome, and systems with near-zero within-class correlation are representative of different latent signals with respect to the target outcome. In this interpretation, systems with negative within-class correlation can be understood to represent complementary latent signals with respect to the particular target on which the performance metric is based. In this case, that target is an ordering of the samples in the dataset that perfectly separates both classes: *an* ordering, because there are inherently many ($N_{C1}!*N_{C2}!/2$) such orderings that have the same AUROC. By calculating the average within-class correlation as described, rather than the global correlation, the target is implicitly taken into account, and the within-class correlation metric then directly quantifies the complementarity of systems with respect to the target.

DIRAC has broad potential implications

There are at least four direct ways in which the results of this study may be useful where system fusion is applicable, i.e., in a domain within which multiple scoring systems can be constructed.

First, the study described here provides strong evidence for the generally accepted but imprecise idea that combining system pairs that are diverse is likely to be a beneficial strategy, and the results presented quantitatively describe the extents to which this intuition applies in practice. This includes a demonstration of the specific utility of fusing negatively correlated systems. These results suggest that a beneficial approach, when constructing classification systems, may be to select component pieces that have a high *a priori* likelihood of having a null or negative within-class correlation (and are therefore complementary). One approach might fuse systems built on unrelated or inversely related subdomains of the problem at hand, for example fusing a system built on gene expression data with one built on categorical environmental variables. Another approach might involve the fusion of two very different statistical approaches built on the same dataset, fusing, for instance, a system that identifies single strong variables (e.g., LASSO) with one that distributes predictive power (e.g., projection methods).

Second, the DIRAC framework provides a direct test of whether a less accurate system can be successfully fused with a more accurate system. Historically, it was known that including a less capable classifier system in a fusion can sometimes boost overall accuracy, but the reasons for this improvement could not be generally described, and the resulting performance was only weakly predictable at best. With the accuracy of an initial system assessed in a test population, the DIRAC framework indicates the combination of within-class correlation and AUROC that is needed in a second system for the fusion of these two systems to outperform either alone. Perhaps more importantly, it establishes hard boundaries below which fusing a second, less accurate system can never possibly help. In this way, the DIRAC framework suggests a targeted, iterative approach to classifier system construction.

Third, the DIRAC framework suggests that estimates of each system's AUROC and the within-class correlation may be obtained separately from each other in space or time, provided the populations in which they are determined are sufficiently similar. This is a potential advantage when attempting to integrate the results of disparate previous studies or when selecting previously tested systems to include in a fusion system under construction.

Fourth, the DIRAC framework implies that any two highly correlated, equally accurate scoring systems have essentially equivalent utility. This means that the easier or cheaper option may be selected without compromising the overall accuracy of the fused system.

Because the DIRAC framework is inherently domain independent, it is expected to be applicable in far-ranging areas. As DIRAC is based only on sample rankings, the only assumption made about the scoring systems to be fused is that they are monotonically increasing functions. This decouples the DIRAC ***ACCURACY/DIVERSITY*** relationship from all other aspects of the scoring systems, such as their absolute magnitudes, their distributions, or their outlier and error behavior, and results in system fusion behavior that is truly domain independent. Applications of the DIRAC framework are thus likely to be found in wide-ranging areas such as clinical biomarker development/personalized medicine (e.g., to determine whether combinations of specific markers can be beneficial, to optimize information gain relative to cost, and integrate multiple information sources such as clinical chemistry and clinical phenotypes), clinical trial enrollment (e.g., maximize enrollment of informative subjects), insurance pricing (e.g., to leverage distinct information streams about potential risks), portfolio management (e.g., maximizing information gain from a variety of market and pricing models, or effectively combining predictors of potential gain with predictors of potential risk), and sensor network optimization.

Score fusions versus rank fusions

We now revisit the differences observed between scoring system fusions and ranking system fusions. As noted above, ranking system fusions are more predictable than their scoring system counterparts, especially at lower sample sizes (N). Despite the

loss of information inherent in converting scores to ranks, many classification problems are well represented by sample rankings. These include such "top-N/bottom-N" problems as separating the top and bottom quintiles of a population in terms of disease risk or selecting the ten best performing stocks for a portfolio. Often these types of problems interface with an external constraint. For example, if a researcher only has enough funding to enroll 100 patients in a clinical trial, what is important is selecting the 100 best candidates. Of less importance is the exact numerical difference between the 100th and the 101st candidates. Instead of constructing full predictive systems of classifier score distribution, then establishing a numerical threshold for identifying the "top-N" samples, the ranking system fusion approach, applied as described in the DIRAC framework, suggests a method for system construction using the rankings of the samples directly.

Although ranking system fusions may be more predictable in general, scoring system fusions may have distinct utility in certain situations. Score fusions can have a higher absolute gain in accuracy compared with their ranking system equivalents (cf., Figures 2 and 4). Furthermore, fusions built around scoring systems can enable the creation of discriminative systems with absolute numerical bounds and address situations where a test subject ranking does not convey adequate information (e.g., within a highly unbalanced dataset). We hypothesize that the specifics of score-based fusion approaches are likely to be more domain dependent. It is likely that the additional information separating a ranking system from its scoring system progenitor originates from the interaction of specific score distribution shapes, outlier and error generating processes, and other domain-dependent factors. Making use of the extra information present in the score distributions and combining those approaches into this DIRAC framework is another area of future interest.

The relationship between ***ACCURACY*** and ***DIVERSITY***, widely acknowledged to play a fundamental role in information fusion, has long eluded a precise, quantitative, and general formulation; the DIRAC framework presented in this study introduces such a formulation. From its origin within this specific context, this framework opens multiple avenues for future research into both the mathematical underpinnings and possible extensions of the observations presented in this work, and their practical applications (M.J.S. et al., unpublished data). This being a limited study, it was impossible to empirically explore the full range of possible methods for combining systems and the metrics for measuring ***ACCURACY*** and ***DIVERSITY***. In the DIRAC framework as presented, we focused on a single fusion approach (pairwise averaging), a single accuracy metric (AUROC), and a single diversity metric (Pearson and Spearman correlation for score and rank, respectively). In addition to their specific utility in our work, these metrics are well understood, generally applicable, and popular. The integration of other fusion approaches and performance metrics into the DIRAC framework is a key area for future research. Additionally, our restriction to pairwise system fusions does not preclude the fusions of more than two systems within this framework, but it does restrict such a construction to an iterative series of pairwise fusions, similar in practice to forward stepwise regression. This sequential fusion approach, and an extension to simultaneous fusion of multiple (>2) scoring systems, is another promising area for future work (M.J.S. el al., unpublished data).

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Bruce Kristal (bkristal@bwh.harvard.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The accession number for the imaging data (reduced form only—no identifiers) reported in this paper is Zenodo: https://doi.org/10.5281/zenodo.5711898. The accession number for the simulation and figure generation code used in this paper (which includes the imaging data as a dependency) is Zenodo: https://doi.org/10.5281/zenodo.5711948. Code is freely available for non-commercial use; inquiries concerning commercial use should be addressed to the lead contact, Bruce Kristal (bkristal@bwh.harvard.edu).

Simulations: Scoring systems based on Gaussian and EMG data

To simulate C_1 and C_2 data using a Gaussian density function, four parameters must first be chosen: the means and standard deviations of the C_1 and C_2 distribution. To enforce some degree of separation between the component distributions, the mean of the C_1 component was sampled at random from the range -10 to 0 and the mean of the C_2 component was sampled at random from the range 0–10. The standard deviations of both components were then sampled uniformly and at random from the range 1–20. For EMGs, a separate exponential distribution was parameterized (separately for both the C_1 and C_2 datasets) by selecting the exponential mean uniformly and at random from the range 0–1,000. To sample exponentially modified data, a sample is first drawn from the Gaussian distribution, then a sample is drawn from the corresponding exponential distribution, and these samples are added together. With the sampling distributions parameterized, an equal number of samples were then drawn from the C_1 and C_2 distributions and pooled. This created a single instance of a simulated C_1 versus C_2 scoring system. To explore the effect of the number of samples drawn (simulating the N of the experiment), this process was repeated many times, with the sample number ranging from 10 to 600/class. All systems were scaled from 0 to 1. If necessary, one scoring system may be inverted (multiplied by -1.0, then rescaled) to properly orient the class separation they induce (i.e., mean score$[C_1] \leq$ mean score$[C_2]$). Simulations were coded in Haskell (ghc 7.10.3). The process is shown schematically in Figure 7, and we note that these distributions are representative of those observed in our real-world studies (see Figure S18).

Simulations: Matching SS_A to SS_B

Simulated scoring systems do not correspond to multiple measurements of a single set of phenomena. Therefore, the scores that are sampled from the C_1 and C_2 distributions of each of the pair of systems can be explicitly paired in order to influence the apparent correlation of the two scoring systems. When the $C_1\_SS_A$ to $C_1\_SS_B$ and $C_2\_SS_A$ to $C_2\_SS_B$ pairings are done at random, the within-class correlation between the two systems will be minimal (i.e., $r \sim 0$). When the pairing is not random and (for example) low-scoring, mid-scoring, and high-scoring case (and respectively control) samples from SS_A are paired with equivalent low-scoring, mid-scoring, and high-scoring case (respectively control) samples from SS_B, a non-zero level of correlation can be induced. In the extreme case, with the points matched exactly by rank, the correlation will achieve a maximum value. Although the actual maximum value will also be influenced by other properties of the sampling distributions, if the distribution shapes are very similar the maximum correlation will approach 1.0; similar but opposite strategies can be used to build inversely correlated dataset. Simulations were coded in Haskell (ghc 7.10.3). The process is shown schematically in Figure 8.

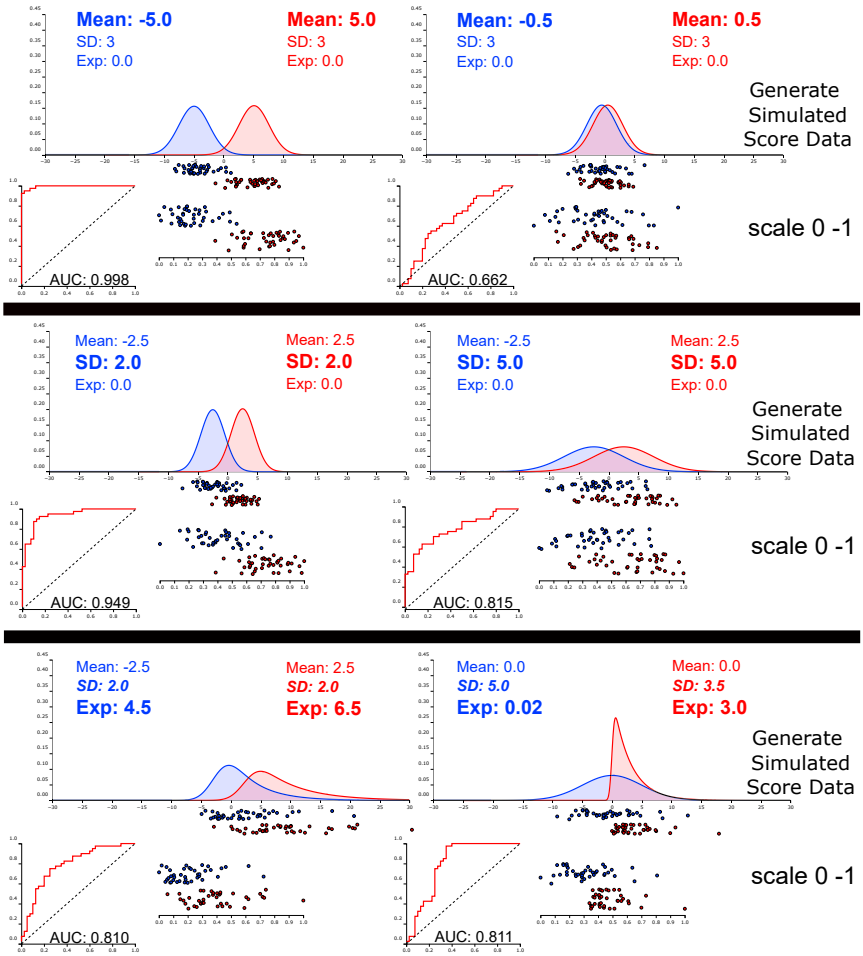


Figure 7. Schematics showing details of parameterization of primary models

Each example shows scoring system pairs generated by Gaussian and exponentially modified Gaussian probability distributions. Each row manipulates a different one of our three main parameters on the classification accuracy (as shown by AUROC plot). The top row shows the effect of altering the C_1 and C_2 Gaussian means. The middle row shows the effect of altering the Gaussian standard deviation (SD). The bottom row shows the influence of the exponential mean parameter. The red and blue points directly under the distributions are raw samples drawn from those distributions. These data are then scaled to fall within the range 0–1; the normalized points are shown next to the respective ROC curves. The main parameter altered is bolded and enlarged. In the bottom right panel, both the Gaussian SD and the exponential was modified (all SDs in that row are shown in bold italic to highlight this).

Fusion

The process of fusion creates a single, new scoring system from two input scoring systems. In this study we examined the simplest type of fusion, which simply averages the value of the two input scoring systems in a pairwise fashion, after each input scoring system has been scaled to fall between a minimum value of 0.0 and a maximum value of 1.0. Ties were stochastically broken. Simulations and fusions were coded in Haskell (ghc 7.10.3). The process is shown schematically in Figure 8.

LOWESS curves

LOWESS curves were created in R version 3.5.2, and were trained on the rank fusion data shown in Figure 4.

Real data: DXA/MRI data from MEC-APS

To test the applicability of the fusion techniques presented in this work to real-world data, we used MRI-based measures of body fat distribution (e.g., liver fat, visceral fat at the L1-L2 vertebral boundary, etc., see Data S1 and Data S2 for full lists) as the ground-truth target variables, and DXA variables that captured general body fat serving as the predictor variables to fuse. These data were drawn from the MEC-APS, in which 1,861 individuals from the Multi-Ethnic Cohort had their body composition measured by DXA and their abdominal fat distribution assessed by MRI between L1 and L5 between 2013 and 2016. This study was approved by the Institutional Review Board. Data from the initial 1,000 participants recruited were used. Details on the MEC itself and the imaging study have been published, but are not critical for the current study.[42]

We divided each of the target MRI variables into quantiles (tertiles, quartiles, and quintiles), and measured how well each DXA variable was able to discriminate the top quantile from the bottom in terms of AUROC. We then fused all possible pairs of DXA predictors, measuring the within-class correlation between the two input scoring systems and the AUROC of the fused system. This fusion performance was then compared with that of the simulated data. The target variables were coded using a dummy variable that assigned the top quantile to class 2, the bottom quantile to class 1, and discarded samples from the other quantiles.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100415.

ACKNOWLEDGMENTS

We would like to thank L. Le Marchand, U. Lim, and K. Monroe for their work related to the MEC datasets presented. B.S.K. acknowledges his family's support, most especially his daughters Rebekah and Bella, and his wife, Linda. The development of this work was supported by grants from the NIH: R01-AG28996 (B.S.K.), U01-ES16048 (B.S.K.), RC1-ES018411 (B.S.K.), R01-AG045713 (B.S.K.), R01-HL132556 (B.S.K.), P01-CA168530 (L.L. Marchand, project 3, director: B.S.K.), R01-HL140335 (B.S.K.).

AUTHOR CONTRIBUTIONS

B.S.K., M.J.S., and D.F.H. have collaborated in this general area and had underlying scientific discussions that set the stage for this work. M.J.S. and B.S.K. conceived the specific project and designed, conducted, and analyzed

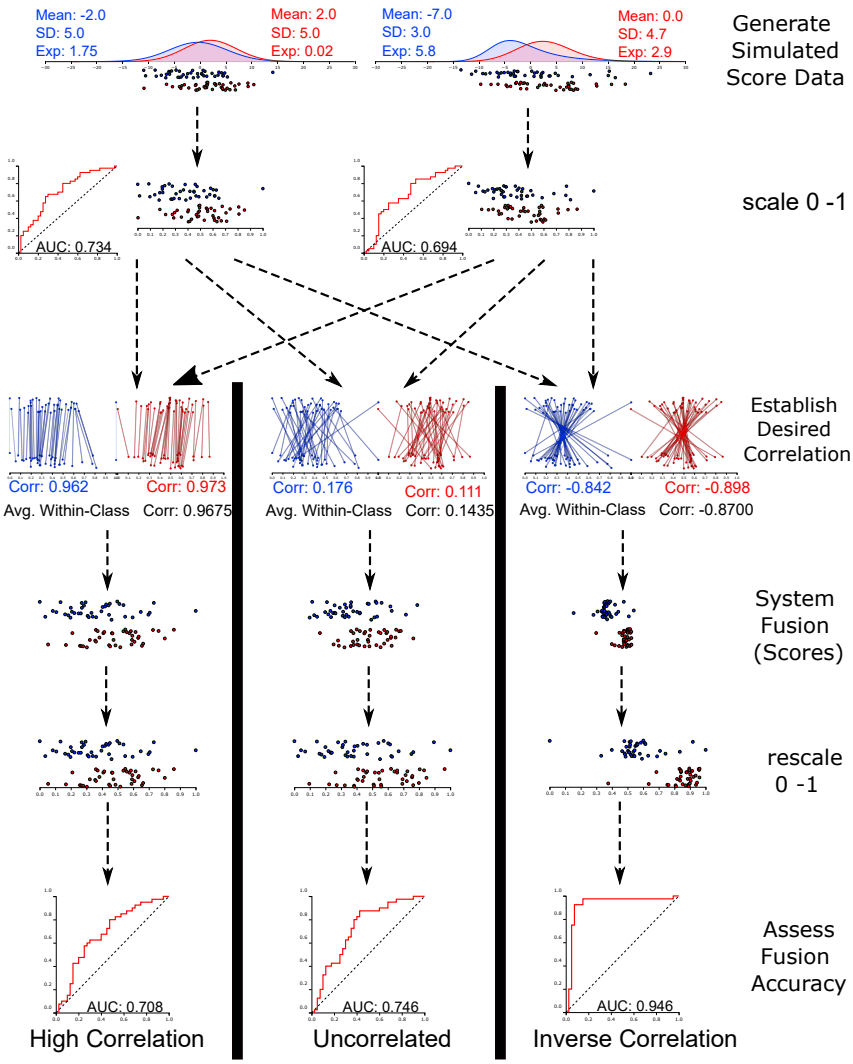


Figure 8. Schematics showing a detailed full representation of pairwise system fusion at the score level

First, two scoring systems are generated and scaled as described in Figure 7. The correlation induction algorithm then pairs the C_1 samples from SS_A and SS_B, and the C_2 samples from SS_A and SS_B, in a way that influences the resulting within-class correlation (bottom three columns separated by black bars). On the left, low-valued C_1 samples from SS_A are preferentially paired with low-valued C_1 samples in SS_B, and this is repeated for C_2, resulting in relatively high correlation. In the center, C_1 samples from SS_A and SS_B are paired randomly, and this is repeated for C_2, resulting in low values of correlation. On the right, low-valued C_1 samples from SS_A are preferentially paired with high-valued C_1 samples in SS_B and vice versa, and this is repeated for C_2, resulting in a negative correlation. Note that, for visual clarity, these show Ns of 40/class; main studies in the report use Ns of 300/class, except when varied as a test (Figure 3).

the "experiments." D.F.H. contributed to scientific discussions of this work, especially with respect to the context of the larger computational field to which this work belongs and the rank/score function, a distinct, complementary method he developed.[30] J.A.S. (DXA), T.E. (MRI), L.R.W. (Statistics), and B.S.K. and M.J.S. (general) were involved in the MEC data generation and integration in this study, along with (from acknowledgments) L.L.M. (PI of MEC P01), U.L., and K.M. M.J.S. and B.S.K. wrote the manuscript, which all authors edited and approved.

DECLARATION OF INTERESTS

M.J.S., D.F.H., and B.S.K. are inventors on a PCT patent application filed by MassGeneral Brigham that covers related technology. The authors declare no other competing interests.

Received: June 10, 2021
Revised: September 20, 2021
Accepted: November 24, 2021
Published: December 22, 2021

REFERENCES

1. de Borda, J.-C. (1781). Mémoire sur les élections au scrutin. Histoire de L'Académie Royale des Sci. *102*, 657–665.
2. Le Marquis de Condorcet. (1785). Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix (Imprimerie Royale).
3. Batallones, A., Sanchez, K., Mott, B., Coffran, C., and Hsu, D.F. (2015). On the combination of two visual cognition systems using combinatorial fusion. Brain Inform. *2*, 21–32.
4. Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G., and Frith, C.D. (2010). Science *329*, 1081–1085.
5. Chung, Y.-S., Hsu, D.F., and Tang, C.Y. (2007). On the diversity-performance relationship for majority voting in classifier ensembles. In 7th International Workshop on Multiple Classifier Systems, M. Haindl, K. Kittler, and F. Roli, eds. (Springer), pp. 407–420.
6. Tumer, K., and Ghosh, J. (1996). Error correlation and error reduction in ensemble classifiers. Connect. Sci. *8*, 385–404.
7. Ruta, D., and Gabrys, B. (2000). An overview of classifier fusion methods. Comput. Inf. Syst. *7*, 1–10.
8. Xu, L., Krzyzak, A., and Suen, C.Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans. Syst. Man Cybern. *22*, 418–435.
9. Clemen, R.T., and Winkler, R.L. (1999). Combining probability distributions from experts in risk analysis. Risk Anal. *19*, 187–203.

10. Jaafar, H., and Ramli, D.A. (2013). A review of multibiometric system with fusion strategies and weighting factor. Int. J. Comput. Sci. Eng. *2*, 158–165.

11. Castanedo, F. (2013). A review of data fusion techniques. Sci. World J. *2013*, 704504.

12. Hsu, D.F., Chung, Y.-S., and Kristal, B.S. (2006). Combinatorial fusion analysis: methods and practices of combining multiple scoring systems. In Advanced Data Mining Technologies in Boinformatics, H.H. Hsu, ed. (IGI Global), pp. 32–62. 10.4018/978-1-59140-863-5.

13. Hall, D.L., and Jordan, J.M. (2010). Human-Centered Information Fusion (Artech House).

14. Ng, K.B., and Kantor, P.B. (2000). Predicting the effectiveness of naive data fusion on the basis of system characteristics. J. Assoc. Inf. Sci. Technol. *51*, 1177–1189.

15. Kittler, J., and Alkoot, F.M. (2003). Sum versus vote fusion in multiple classifier systems. IEEE Trans. Pattern Anal. Mach. Intell. *25*, 110–115.

16. Crisman, K.-D. (2014). The Borda count, the Kemeny rule, and the permutahedron, in the mathematics of decisions, elections, and games. Am. Math. Soc. *624*, 101–134.

17. Sun, S.B., Zhang, Z.H., Dong, X.L., Zhang, H.R., Li, T.J., Zhang, L., and Min, F. (2017). Integrating Triangle and Jaccard similarities for recommendation. PLoS One *12*, e0183570.

18. Chung, Y.-S., Hsu, D.F., Liu, C.-Y., and Tang, C.-Y. (2010). Performance evaluation of classifier ensembles in terms of diversity and performance of individual systems. Int. J. Pervasive Comput. Commun. *6*, 373–403.

19. Tang, E.K., Suganthan, P.N., and Yao, X. (2006). An analysis of diversity measures. Machine Learn. *65*, 247–271.

20. Kuncheva, L.I. (2003). That elusive diversity in classifier ensembles. In First Iberian Conference on Pattern Recognition and Image Analysis, F.J. Perales, A.J.C. Campilho, N. Pérez de la Blanca, and A. Sanfeliu, eds. (Springer), pp. 1126–1138.

21. Kuncheva, L.I. (2004). Combining Pattern Classifiers: Methods and Algorithms (John Wiley & Sons).

22. Guo, B., Nixon, M.S., and Damarla, T. (2012). Improving acoustic vehicle classification by information fusion. Pattern Anal. Appl. *15*, 29–43.

23. Kantor, P.B., Ng, K.B., and Hull, D. (1998). Comparison of System Using Pairs-Out-Of-Order. Technical Report (National Institute of Standards and Technology).

24. Wolpert, D.H., and Macready, W.G. (2005). Coevolutionary free lunches. IEEE Trans. Evol. Comput. *9*, 721–735.

25. Forster, M.R. (2005). Notice: No Free Lunches for Anyone, Bayesians Included (Department of Philosophy, University of Wisconsin–Madison).

26. Ho, Y.-C., and Pepyne, D.L. (2001). Simple explanation of the no free lunch theorem of optimization. In Proceedings of the 40th IEEE Conference on Decision and Control, *5* (IEEE), pp. 4409–4414.

27. Wolpert, D.H., and Macready, W.G. (1997). No free lunch theorems for optimization. IEEE Trans. Evol. Comput. *1*, 67–82.

28. Wolpert, D.H. (1996). The lack of a priori distinctions between learning algorithms. Neural Comput. *8*, 1341–1390.

29. Wu, S. (2007). A geometric probabilistic framework for data fusion in information retrieval. In 10th International Conference on Information Fusion (IEEE), pp. 1–8.

30. Hsu, D.F., and Taksa, I. (2005). Comparing rank and score combination methods for data fusion in information retrieval. Inf. Retrieval *8*, 449–480.

31. Vogt, C.C., and Cottrell, G.W. (1999). Fusion via a linear combination of scores. Inf. retrieval *1*, 151–173.

32. Chen, Y.F., Hsu, K.C., Lin, P.T., Hsu, D.F., Kristal, B.S., and Yang, J.M. (2011). LigSeeSVM: ligand-based virtual screening using support vector machines and data fusion. Int. J. Comput. Biol. Drug Des. *4*, 274–289.

33. Whittle, M., Gillet, V.J., Willett, P., and Loesel, J. (2006). Analysis of data fusion methods in virtual screening: theoretical model. J. Chem. Inf. Model. *46*, 2206–2219.

34. Yang, J.M., Chen, Y.F., Shen, T.W., Kristal, B.S., and Hsu, D.F. (2005). Consensus scoring criteria for improving enrichment in virtual screening. J. Chem. Inf. Model. *4*, 1134–1146.

35. Ernst, M.O. (2010). Decisions made better. Science *329*, 1022–1023.

36. Melnik, O., Vardi, Y., and Zhang, C.-H. (2004). Mixed group ranks: preference and confidence in classifier combination. IEEE Trans. Pattern Anal. Mach. Intell. *26*, 973–981.

37. Bedo, J., and Ong, C.S. (2016). Multivariate spearman's rho for aggregating ranks using copulas. J. Mach. Learn. Res. *17*, 1–30.

38. Lebanon, G., and Lafferty, J. (2002). Cranking: combining rankings using conditional probability models on permutations. ICML *2*, 363–370.

39. Hanley, J.A., and McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology *143*, 29–36.

40. Lim, U., Monroe, K.R., Buchthal, S., Fan, B., Cheng, I., Kristal, B.S., Lampe, J.W., Hullar, M.A., Franke, A.A., Stram, D.O., et al. (2019). Propensity for intra-abdominal and hepatic adiposity varies among ethnic groups. Gastroenterology *156*, 966–975.

41. Lenchik, L., Register, T.C., Hsu, F.C., Lohman, K., Nicklas, B.J., Freedman, B.I., Langefeld, C.D., Carr, J.J., and Bowden, D.W. (2003). Adiponectin as a novel determinant of bone mineral density and visceral fat. Bone *33*, 646–651.

42. Sheu, Y., and Cauley, J.A. (2011). The role of bone marrow and visceral fat on bone metabolism. Curr. Osteoporos. Rep. *9*, 67–75.

43. Thakkar, A., and Chaudhari, K. (2021). Fusion in stock market prediction: a decade survey on the necessity, recent developments, and potential future directions. Inf. Fusion *65*, 95–107.

44. Modak, S.K.S., and Jha, V.K. (2019). Multibiometric fusion strategy and its applications: a review. Inf. Fusion *49*, 174–204.

45. Fierrez, J., Morales, A., Vera-Rodriguez, R., and Camacho, D. (2018). Multiple classifiers in biometrics. Part 1: fundamentals and review. Inf. Fusion *44*, 57–64.

46. Barak, S., Arjmand, A., and Ortobelli, S. (2017). Fusion of multiple diverse predictors in stock market. Inf. Fusion *36*, 90–102.

47. Arevalo, J., Solorio, T., Montes-y Gomez, M., and Gonzalez, F.A. (2017). Gated multimodal units for information fusion. arXiv 1702.01992.

48. Kim, J., Koh, J., Kim, Y., Choi, J., Hwang, Y., and Choi, J.W. (2018). Robust deep multi-modal learning based on gated information fusion network. In 14th Asian Conference on Computer Vision, C.V. Jawahar, H. Li, G. Mori, and K. Schindler, eds. (Springer), pp. 90–106.

49. Honarmandi, P., Duong, T.C., Ghoreishi, S.F., Allaire, D., and Arroyave, R. (2019). Bayesian uncertainty quantification and information fusion in calphad-based thermodynamic modeling. Acta Mater. *164*, 636–647.

50. Chen, C., Zhang, Q., Ma, Q., and Yu, B. (2019). LightGBM-PPI: Predicting protein-protein interactions through light with multi-information fusion. Chemometr. Intell. Lab. Syst. *191*, 54–64.

51. Yunyun, Z., Xiangke, W., Weiwei, K., and Lincheng, S. (2017). Information fusion analysis of multi-UAV system based on information geometry. In 36th Chinese Control Conference (CCC) (IEEE), pp. 8491–8496.

52. Al Bitar, N., and Gavrilov, A. (2019). Comparative analysis of fusion algorithms in a loosely-coupled integrated navigation system on the basis of real data processing. Gyroscopy Navig. *10*, 231–244.

53. Balazs, J.A., and Velasquez, J.D. (2016). Opinion mining and information fusion: a survey. Inf. Fusion *27*, 95–110.

54. Benham, R., and Culpepper, J.S. (2017). Risk-reward trade-offs in rank fusion. In Proceedings of the 22nd Australasian Document Computing Symposium (Association for Computing Machinery), pp. 1–8.

55. L. Snidaro, J. García, J. Llinas, and E. Blasch, eds. (2016). Context-Enhanced Information Fusion. Boosting Real-World Performance with Domain Knowledge (Springer).

56. Raz, A.K., Kenley, C.R., and DeLaurentis, D.A. (2017). A system-of-systems perspective for information fusion system design and evaluation. Inf. Fusion *35*, 148–165.

57. Snidaro, L., Garcia, J., Llinas, J., and Blasch, E. (2019). Recent trends in context exploitation for information fusion and AI. AI Mag. *40*, 14–27.

58. Ali, F., El-Sappagh, S., Islam, S.R., Kwak, D., Ali, A., Imran, M., and Kwak, K.-S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Inf. Fusion *63*, 208–222.

59. Zhou, L., Zhang, C., Qiu, Z., and He, Y. (2020). Information fusion of emerging non-destructive analytical techniques for food quality authentication: a survey. Trac Trends Anal. Chem. *127*, 115901.

60. Zhang, L., Pan, Y., Wu, X., and Skibniewski, M.J. (2021). Information fusion. Artificial Intelligence in Construction Engineering and Management (Springer), pp. 95–124.

61. Liu, Z., Wu, J., Fu, L., Majeed, Y., Feng, Y., Li, R., and Cui, Y. (2019). Improved kiwifruit detection using pre-trained VGG16 with RGB and NIR information fusion. IEEE Access *8*, 2327–2336.

62. Xiao, Y., Li, C., Song, L., Yang, J., and Su, J. (2021). A multidimensional information fusion-based matching decision method for manufacturing service resource. IEEE Access *9*, 39839–39851.

63. Xiao, Y., Xue, J., Zhang, L., Wang, Y., and Li, M. (2021). Misalignment fault diagnosis for wind turbines based on information fusion. Entropy *23*, 243.

64. Cruz, R.M., Sabourin, R., and Cavalcanti, G.D. (2018). Dynamic classifier selection: recent advances and perspectives. Inf. Fusion *41*, 195–216.

65. Wang, H., Ding, Y., Tang, J., and Guo, F. (2020). Identification of membrane protein types via multivariate information fusion with Hilbert-Schmidt independence criterion. Neurocomputing *383*, 257–269.

66. Wang, W., Zhang, Z., Qi, S., Shen, J., Pang, Y., and Shao, L. (2019). Learning compositional neural information fusion for human parsing. In Proceedings of the IEEE/CVF International Conference on Computer Vision (IEEE), pp. 5703–5713.

67. Perdikaris, P., Raissi, M., Damianou, A., Lawrence, N.D., and Karniadakis, G.E. (2017). Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. Proc. R. Soc. A Math. Phys. Eng. Sci. *473*, 20160751.

68. Gao, L., Qi, L., Chen, E., and Guan, L. (2017). Discriminative multiple canonical correlation analysis for information fusion. IEEE Trans. Image Process. *27*, 1951–1965.

69. Brown, G. (2009). An information theoretic perspective on multiple classifier systems. In MCS: International Workshop on Multiple Classifier Systems, J.A. Benediktsson, J. Kittler, and F. Roli, eds. (Springer), pp. 344–353.

70. Ibraev, U., Ng, K.B., and Kantor, P.B. (2002). Exploration of a geometric model of data fusion. Proc. Assoc. Inf. Sci. Technol. *39*, 124–129.

71. Wu, S., and Crestani, F. (2015). A geometric framework for data fusion in information retrieval. Inf. Syst. *50*, 20–35.

72. Muhammad, G., Alshehri, F., Karray, F., El Saddik, A., Alsulaiman, M., and Falk, T.H. (2021). A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. Inf. Fusion *76*, 355–375.

73. Simjanoska, M., Kochev, S., Tanevski, J., Bogdanova, A.M., Papa, G., and Eftimov, T. (2020). Multi-level information fusion for learning a blood pressure predictive model using sensor data. Inf. Fusion *58*, 24–39.

74. Zhang, C., Yang, Z., He, X., and Deng, L. (2020). Multimodal intelligence: representation learning, information fusion, and applications. IEEE J. Selected Top. Signal Process. *14*, 478–493.

75. Gupta, K., Walia, G.S., and Sharma, K. (2020). Quality based adaptive score fusion approach for multimodal biometric system. Appl. Intell. *50*, 1086–1099.

76. Lumini, A., and Nanni, L. (2017). Overview of the combination of biometric matchers. Inf. Fusion *33*, 71–85.