

Revealing the Reality: A Comprehensive Framework for Evaluating and Selecting Score-Based Predictive Models

Kenji Tanaka^{1*}, Yui Kobayashi¹, Takeshi Nakamura²

¹ Department of Biostatistics and Predictive Modeling, Faculty of Medicine, University of Tokyo, Tokyo, Japan.

² Department of Health Informatics and Model Evaluation, Faculty of Medicine, Kyoto University, Kyoto, Japan.

*E-mail ✉ tanaka.kenji@med.u-tokyo.ac.jp

Abstract

Choosing appropriate machine learning (ML) models in biomedical research typically depends on overall performance indicators. However, when these indicators yield very similar results across candidate models, determining the best option for practical clinical use becomes difficult. We developed an innovative combined approach that integrates visual examination of the Model Scoring Distribution Analysis (MSDA) with a newly introduced scoring measure (MSDscore). This technique was integrated into the Digital Phenomics platform via the MSDanalyser tool. It was evaluated by creating and assessing 27 predictive models for prognosis in breast, lung, and renal cancers. The method facilitated an in-depth review of the distributions of true positives, false positives, true negatives, and false negatives throughout the scoring range. This uncovered localized performance trends that standard global metrics often miss. While traditional metrics showed only slight differences among the models, the MSDA framework highlighted notable variations in behavior across specific score regions, enabling clearer differentiation. Combining this composite framework with conventional performance measures offers a valuable supplementary perspective, providing greater insight into model selection in clinical and biomedical contexts.

Keywords: Modelling, Biomedical, Machine-learning, Cancer, Prognostics

Introduction

The incorporation of machine learning (ML) techniques into biomedical studies has transformed predictive model development, especially for applications such as disease forecasting, patient grouping, and predicting responses to therapies [1-5]. These models commonly use complex, high-dimensional datasets—such as genomic, transcriptomic, and other omics data—to produce risk scores or probability estimates that guide clinical decisions. In line with regulatory requirements for medical use, most of these models calculate a continuous

score by applying mathematical formulas to a core set of biomarkers (typically gene or protein expression levels) [6, 7]. Determining binary outcomes (positive/negative) for diagnosis or prognosis generally involves applying threshold values derived from the available data [6, 7].

A key phase in implementing these models is selecting the optimal option from several [1, 8]. Conventionally, this decision is based on evaluation metrics such as the area under the receiver operating characteristic (ROC) curve, accuracy, sensitivity, and specificity [9-12]. Although these indicators offer useful overviews of model effectiveness, they may fall short in separating closely performing models, particularly when results across the metrics are nearly identical.

In real-world clinical environments, incorrect classifications can have substantial consequences. For example, in cancer care, adopting a less effective prognostic model might result in misguided therapeutic choices, ultimately harming patient outcomes [10, 13,

Access this article online

<https://smerpub.com/>

Received: 28 August 2023; Accepted: 27 November 2023

Copyright CC BY-NC-SA 4.0

How to cite this article: Tanaka K, Kobayashi Y, Nakamura T. Revealing the Reality: A Comprehensive Framework for Evaluating and Selecting Score-Based Predictive Models. J Med Sci Interdiscip Res. 2023;3(2):159-70. <https://doi.org/10.51847/0sIEf5eE4>

14]. This highlights the need for more refined assessment strategies that can detect subtle distinctions in model behavior beyond conventional metrics.

Various strategies have been suggested to tackle this issue. Visual analysis platforms such as Manifold and HypoML support model understanding and comparison via dynamic graphics [15, 16]. Statistical techniques such as Decision Curve Analysis (DCA) assess the practical clinical value of models by calculating net benefits across a range of probability thresholds [17]. Despite their contributions to deeper model insights, these approaches lack a unified scoring mechanism that fully accounts for the overall patterns of true positives, false positives, true negatives, and false negatives—factors essential to actual deployment. Additionally, many current tools are not tailored for threshold-dependent classifiers that produce continuous prediction scores, a common feature in biomedical ML.

To fill this gap, we introduce a new evaluation framework that combines visual and quantitative elements, specifically designed for score-generating models. This method improves model selection by providing graphical displays of distributions for true positives, false positives, true negatives, and false negatives, along with a composite metric that highlights performance details missed by standard indicators. The graphical aspect helps users intuitively grasp the balance among various error types, while the composite score provides a numerical summary that integrates these elements.

In earlier work, we built an AI-powered system, the Digital Phenomics Platform, that supports automated generation and assessment of predictive models via an autoML workflow [18]. This platform provides access to carefully prepared public tumor biopsy transcriptomic datasets for model building. Previously, it incorporated standard performance metrics for model assessment. In this study, we introduced our new approach, the MSDAnalyser tool, and incorporated it directly into the Digital Phenomics Platform. To illustrate its value, we tested the framework on survival-prognosis models trained on breast, lung, and renal cancer datasets. This work benchmarks the effectiveness of the new composite scoring methodology for selecting score-based models, using cancer prognostic examples. We examined how visual and scoring-based assessments compare to and complement traditional performance metrics, revealing insights not visible through standard approaches alone.

Materials and Methods

Cancer transcriptomics datasets

The datasets for model construction were compiled from publicly available data in The Cancer Genome Atlas (TCGA), focusing on tumor biopsy specimens from individuals with breast, lung, and renal cancers. We obtained gene expression profiles from the 2021 release of the Human Protein Atlas, which covers mRNA levels for 200 genes in 1075 de-identified patient samples [19, 20]. These were already normalized and reported as fragments per kilobase of transcript per million mapped reads (FPKM). To ensure consistency with prior modeling work on the Digital Phenomics platform [18], we followed the established curation process. In particular, we focused on 58 genes linked to critical signaling pathways involved in epithelial-to-mesenchymal transition (EMT), a key mechanism in cancer metastasis and invasion [21]. Survival data were converted into binary prognostic categories: good prognosis for patients surviving beyond five years after diagnosis, and poor prognosis for those who passed away within two years.

The curated transcriptomic datasets for the three cancer types were as follows. Breast cancer (BRCA) included 239 samples (40 poor prognosis, 199 good prognosis). Lung cancers (LUSC and LUAD combined) comprised 325 samples (231 poor prognosis, 94 good prognosis). Renal cancer (including KICH, KIRC, and KIRP subtypes) consisted of 318 samples (108 with poor prognosis, 210 with good prognosis). For each cancer, gene expression data were organized in CSV format, with gene names in the first column and FPKM values for each patient in subsequent columns. Separate metadata files linked each sample to its prognostic label. These datasets are available through the Digital Phenomics platform (<https://digitalphenomics.com>, last accessed on 10 May 2025).

Generation of predictive models

Models were generated using the Digital Phenomics platform (<https://digitalphenomics.com>, accessed on 10 April 2025), user interface version 0.22. Within this environment, we used O2Pmgen version 1.1, a proprietary AutoML system from Bioenhancer Systems LTD, which streamlines the full data selection, training, optimization, and validation pipeline. The tool applies a supervised learning approach driven by a genetic/evolutionary algorithm to discover custom

biomarker sets that optimize sensitivity, specificity, and AUC [18]. For each model, a random subset of the data (always less than 50%) served as the training set. In contrast, the remaining data was reserved for independent testing to replicate practical validation conditions. The process incorporates biomarker pattern recognition that simulates detection of gene up-regulation, down-regulation, and binary activation/inhibition signals, relying on non-parametric statistical cutoffs for reliability [18]. Models were optimized to a specified upper bound on the false-positive rate. The final models follow a general scoring formula for phenotype classification Eq. (1) [18].

$$CS = \sum_i^n 100 \frac{W_i(-P_i + N_i)}{P_i + N_i} \quad (1)$$

Where:

P_i represents the absolute difference between the estimated median value of biomarker i in the phenotype-positive group and the corresponding biomarker value measured in the unknown sample.

N_i denotes the absolute difference between the estimated median value of biomarker i in the control group (phenotype-negative group) and the biomarker value observed in the unknown sample.

W_i denotes the enrichment score for biomarker i in the phenotype-positive group, and n denotes the total number of biomarkers included in the predictive model.

The AI system in the O2Pmgen tool determines the optimal threshold for using the resulting CS equation in phenotype classification (yes/no) [18]. As a result, a model produced by this method classifies a phenotype as yes (positive) when the score exceeds the calculated cutoff and as no (negative) otherwise.

Classical model performance metrics

Performance evaluation of the developed models relied on analytical tools integrated into the Digital Phenomics Platform version 0.22. Receiver Operating Characteristic (ROC) curves were produced with the ROCplot tool (version 1.1). This tool visualizes the relationship between sensitivity and 1-specificity for varying threshold levels on user-selected datasets [11]. It performed a comprehensive sweep across threshold values using a 10,000-step resolution throughout the scoring range. ROC curves generated by ROCplot employ a non-parametric approach to illustrate the

model's capacity to differentiate between phenotypic classes across every possible classification threshold [12, 18]. Additionally, we employed the O2Pmgen tool features within the Digital Phenomics platform to calculate the following indicators: optimal sensitivity (Eq. (2)), optimal specificity (Eq. (3)), optimal accuracy (Eq. (4)), and the area under the ROC curve (AUC).

$$\begin{aligned} \text{Sensitivity (\%)} & \quad (2) \\ &= \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} * 100 \end{aligned}$$

$$\begin{aligned} \text{Specificity (\%)} & \quad (3) \\ &= \frac{\text{True Positives}}{\text{True Negatives} + \text{False Positives}} * 100 \end{aligned}$$

$$\begin{aligned} \text{Accuracy (\%)} &= \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Observations}} \\ & \quad * 100 \end{aligned} \quad (4)$$

The AUC calculation in ROCplot used a 5% performance interval window to improve precision in area estimation via the Riemann sum method [18]. We computed the AUC, along with other evaluation measures, following recommended practices for diagnostic models in biomedical fields, based on the framework proposed by Dankers *et al.* [12] and incorporated into both ROCplot and O2Pmgen. Every performance indicator and the associated ROC curves were exported as CSV files directly from the Digital Phenomics platform's interface tools for additional analysis. ROC curve images were saved as PNGs via the platform's export capabilities.

Model scoring distribution analysis (MSDA)

The MSDA technique investigates how true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are distributed across the full range of scores output by a predictive model during phenotype classification. It examines the spectrum of scores produced by the mathematical scoring function used for classifying the phenotype after applying the determined cutoff threshold. This evaluation utilizes scores from the full set of samples reserved for model validation. The entire score range is partitioned into ten uniform intervals (bins) progressing from lowest to highest values. For every interval, the counts of TP, TN, FP, and FN are tallied. This process reveals the pattern of classification results along the scoring continuum and tests whether higher scores indeed indicate a greater probability of the target phenotype. The choice of ten bins offers a practical

compromise between sufficient detail and interpretative clarity. This setup applies decimal scaling to the scores, making it easier to contrast different models, as with typical normalization techniques. The bin structure highlights particular segments of the score range with unique predictive characteristics. Thus, individual bins serve as localized indicators of model reliability, exposing areas of high confidence or potential weakness. An illustrative MSDA plot was produced as a visual aid for comparing models, proving especially useful when standard global measures such as AUC, sensitivity, specificity, and accuracy yield very close results. The graphic includes a stacked bar chart showing the distribution of TP, TN, FP, and FN counts within each scoring bin, paired with a line plot showing the overall sample count per bin (**Figure 1**).

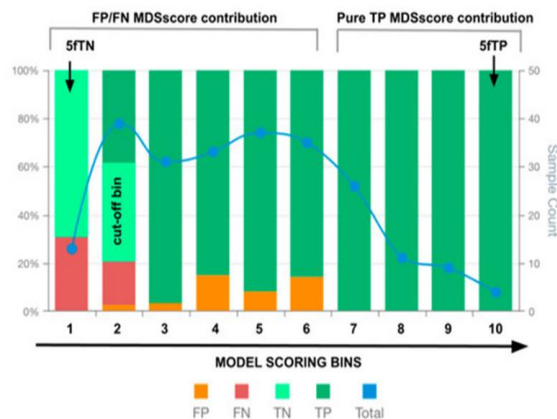


Figure 1. Representative illustration of an MSDA plot.

The plot uses distinct colors to indicate false positives (FP), false negatives (FN), true positives (TP), true negatives (TN), and the overall number of cases (Total) in each score bin, as shown in the accompanying legend. Percentages of TP, TN, FP, and FN within individual bins are plotted against the left vertical axis, while the right vertical axis shows the total observation counts for each bin. The horizontal axis displays the binned intervals of model scores for phenotype classification. Each bin label on the x-axis reflects the upper boundary of that interval (for example, 0–1, 1–2, 2–3, 3–4, and so on). Key parts of the MSD score (Eq. (5)) appear at the top of the figure. Arrows point to the 5fTN and 5fTP terms from Eq. (5), with grey brackets denoting the specific bin regions corresponding to the remaining components of the equation.

To enhance visual evaluation, we introduced a dedicated scoring function, the MSDscore. This metric quantifies the overall quality of how true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are distributed throughout the score bins (Eq. (5)). The score is determined in the following manner:

$$\text{MSDscore} = 5f\text{TN}_{B1} + 5f\text{TP}_{B10} + \sum_i^n 10K_i - \sum_j^m \frac{(FP_j + FN_j)}{\text{Total}_j} \quad (5)$$

Where

- $f\text{TN}_{B1}$ represents the proportion of true negatives in the initial score bin (B1), reflecting the model's accuracy in correctly classifying negative cases at the lowest scores.
 - $f\text{TP}_{B10}$ denotes the proportion of true positives in the highest score bin (B10), indicating how reliably elevated scores identify positive cases.
 - K_i corresponds to the number of samples in bin i located in a marginal zone spanning n bins that contain exclusively correct classifications (no false positives or false negatives). These areas signify regions of strong local model performance.
 - FP_j and FN_j refer to the counts of false positives and false negatives, respectively, in bin j across $m = 9$ bins (all bins except the one holding the model's classification cutoff). This component penalizes prediction errors distributed throughout the scoring range outside the cutoff area.
 - Total_j is the total number of samples present in bin j .
- Eq.(5) was formulated to reward strong performance at the score extremes while penalizing regions prone to misclassification, providing an additional layer of evaluation alongside conventional global metrics. We applied a multiplier of 10 to K_i to assign substantially greater emphasis (one order of magnitude) to pure true-positive regions in high-scoring areas. This rewards models exhibiting excellent specificity in those segments. Coefficients of 5 were used for $f\text{TN}_{B1}$ and $f\text{TP}_{B10}$ to grant comparable but slightly lower extra weighting (half an order of magnitude) to the distribution tails. These extremes merit emphasis because, under a normal distribution assumption, false positives and false negatives are anticipated to occur far less frequently in the outer regions. The MSDA methodology was implemented using the MSDanalyser tool, built with Python 3.10 and incorporated into the Digital Phenomics

Platform. Its visualization components were coded in JavaScript and seamlessly embedded within the platform's toolkit.

Data manipulation and analysis

Further investigation into the relationships between traditional performance indicators and the MSDscore involved supplementary statistical analyses and graphical comparisons conducted outside the Digital Phenomics Platform. These steps took place in a Jupyter Notebook environment running Python 3.10. The matplotlib library was utilized to create various plots, facilitating visual review of the spread and associations among metrics including AUC, sensitivity, specificity, and MSDscore. Data processing was handled using the pandas and numpy libraries, enabling efficient data structuring, model selection, and performance-based ranking. Pearson correlation coefficients were obtained via NumPy's `corrcoef` function, while linear trend lines were generated using NumPy's `polyfit` function for clearer visualization of relationships.

Results and Discussion

We introduced a novel performance indicator (MSDscore) and a new graphical method (MSDA plot) to support model assessment by analyzing the spread of type I and type II errors across the predictive score range. This overall strategy is termed the MSDA methodology. We used this approach to select optimal models from three sets of cancer prognostic predictors (breast, lung, and renal cancers) alongside conventional evaluation metrics.

Leveraging an AI-based modeling workflow on tumor transcriptomic data, we produced 9 predictive models for each cancer type (breast, lung, and renal), yielding 27 models in total. The Receiver Operating Characteristic (ROC) curves for these models indicated generally solid predictive capability (**Figure 2**). Substantial overlap was observed among the ROC curves for models within the same cancer type, suggesting that no single model stood out markedly from the others based on visual inspection of predictive strength.

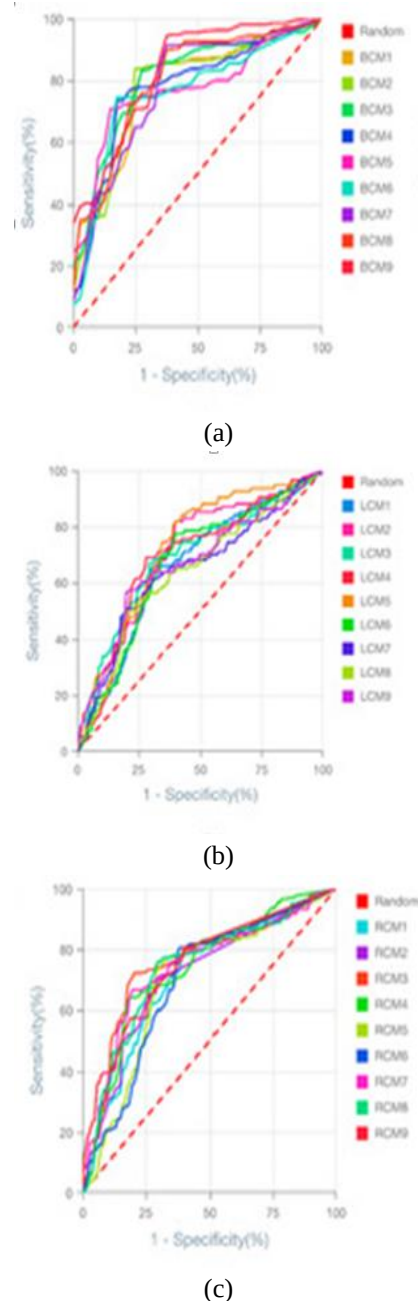


Figure 2. ROC curves generated for the complete set of 27 predictive models developed for cancer prognosis. (a) ROC curves corresponding to breast cancer prognostic models. (b) ROC curves corresponding to lung cancer prognostic models. (c) ROC curves corresponding to renal cancer prognostic models. Individual model labels are provided in the legends. A dashed red line indicates the performance level expected from random guessing, while solid lines show the actual ROC curves for each model.

AUC results fell in the ranges of 75–83% for breast cancer, 63–78% for lung cancer, and 64–75% for renal cancer. Accuracy figures varied from 73–89% (breast), 65–73% (lung), and 70–74% (renal). In comparison, MSDscore values spanned much wider ranges: 51.9–507.7 for breast cancer models, 1.5–15.4 for lung cancer, and 3.0–65.1 for renal cancer. These expanded ranges highlighted the new metric's enhanced ability to effectively distinguish models, both within each cancer category and between different cancer types.

The MSDscore showed moderate Pearson correlations with AUC and accuracy (Figure 3b), suggesting it can serve as a reasonable proxy for overall predictive strength. On the other hand, it maintained only weak positive associations ($r < 0.2$) with sensitivity and specificity, confirming its value as an independent, supplementary tool for model evaluation.

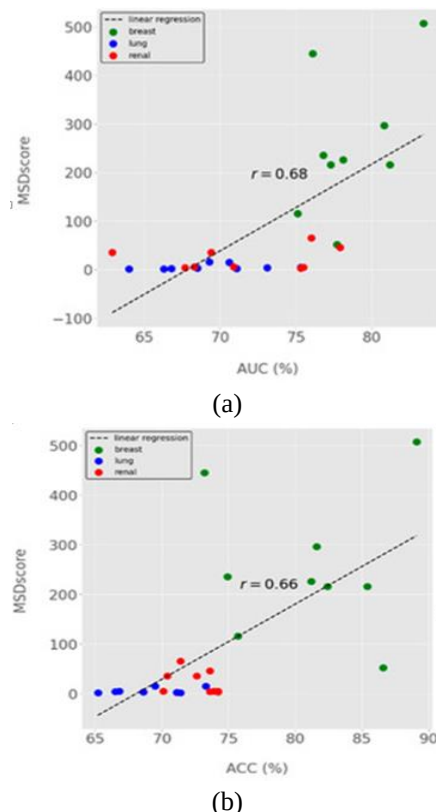


Figure 3. Relationship between MSDscores and conventional performance indicators used for model selection: (a) MSDscores plotted against ROC Area-Under-the-Curve (AUC); (b) MSDscores plotted against accuracy.

Each dot corresponds to the calculated metrics for the 27 models developed for breast, lung, and renal cancer prognosis. Dots of varying colors denote results from different cancer types, as explained in the legend. Dashed lines illustrate the linear regression trend, and the Pearson correlation coefficients (r) are displayed near the center of each panel (a,b) adjacent to the trendlines.

Breast cancer models (BCM) evaluation

Analysis of breast cancer prognostic models indicated that BCM9 was the top performer under standard selection criteria, such as AUC and accuracy. BCM9 also ranked highest under the MSDA methodology, attaining an MSDscore of 507.5—a value roughly ten times that of the other models. Among all breast cancer models, BCM9 had the highest sensitivity (94.5%) but the lowest specificity (62.5%). Its MSD plot showed four high-score bins at the upper end of the range, where true positives were clearly separated from false positives, supported by a large number of samples (Figure 4). This pattern suggests that BCM9 scores exceeding 118.9 are unlikely to yield false positives. Conversely, false negatives were primarily clustered in the first two score bins, indicating that the model tends to generate approximately 30% false negatives for scores up to 36.4. The MSD plot also showed a well-balanced spread of false positives at relatively low rates across the second and sixth bins. Furthermore, the distribution confirmed that each bin contained sufficient observations to reliably estimate the proportions of true and false positives and negatives.

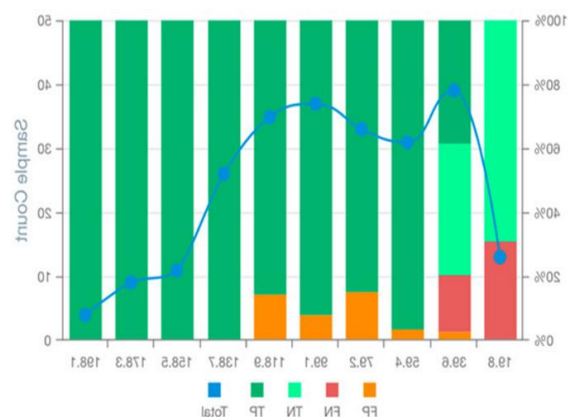


Figure 4. MSD plot for the specialized breast cancer prognostic model BCM9.

The chart uses distinct colors to mark false positives (FP), false negatives (FN), true positives (TP), true negatives (TN), and the overall count of observations (Total) for

each score bin, as indicated in the legend. Percentages of TP, TN, FP, and FN for each bin are shown on the left vertical axis, while the right vertical axis indicates the total observation numbers per bin. The values labeled on the bins correspond to the upper boundary of each interval (e.g., 0–19.8, 19.8–39.6, 39.7–59.4, ...).

Comparative review of the complete collection of MSD plots confirmed that all alternative models fell short of BCM9 in one or several key areas: (1) having fewer score bins entirely without false positives; (2) possessing fewer samples to support those clean high-performance regions; (3) displaying a smaller share of true negatives in the lowest scoring zone; (4) showing more irregular patterns of false positives and negatives surrounding the decision threshold.

Lung cancer model (LCM) evaluation

Evaluation of the lung cancer models revealed that LCM5 achieved the highest AUC (75.3%), while LCM9 achieved the highest accuracy (73.3%). When assessed using the MSDscore, however, LCM9 ranked first (15.4), followed by LCM3 (15.1). Both scores were three to ten times larger than those obtained by the other models, including LCM5 (MSDscore 4.5, third position). Corresponding sensitivity figures were 56% (LCM9), 67% (LCM3), and 80% (LCM5); specificity values stood at 80%, 71%, and 61%, respectively.

MSD plots for these models exposed clear variations in the placement of false positives and false negatives throughout the scoring intervals. Only LCM9 and LCM3 showed zones at the upper score limit completely devoid of false positives (**Figures 5a, 5c**). In detail, LCM3 remained free of false positives beyond 118.8, and LCM9 beyond 163.4. By comparison, LCM5 showed a lingering presence of false positives that extended into elevated scores (though gradually diminishing), a drawback not seen in the other two (**Figure 5b**). Because LCM3, LCM5, and LCM9 exhibited similar rates of true negatives versus false negatives in the initial three bins, discrepancies observed in the later bins carry greater weight when choosing between models. That said, the limited sample sizes in the final two bins for all models constrain the confidence one can place in these trends.

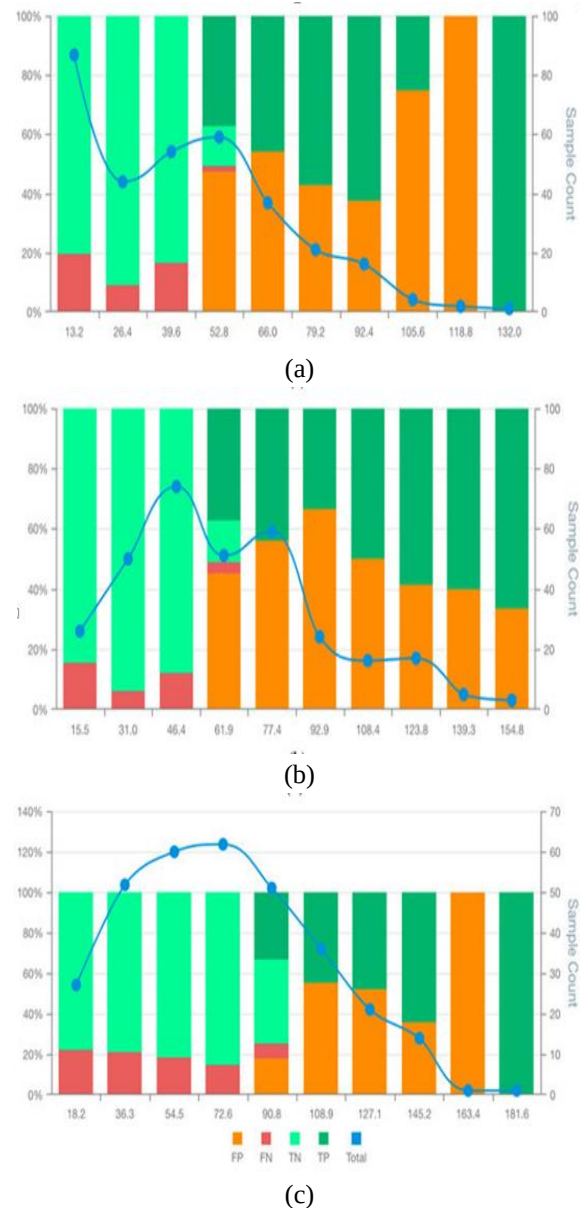
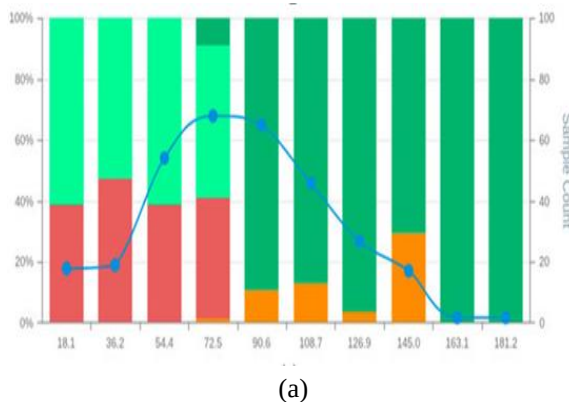


Figure 5. Multiple score distribution analysis (MSDA) plots illustrating the performance of the customized lung cancer prognostic models: LCM3 (a), LCM5 (b), and LCM9 (c).

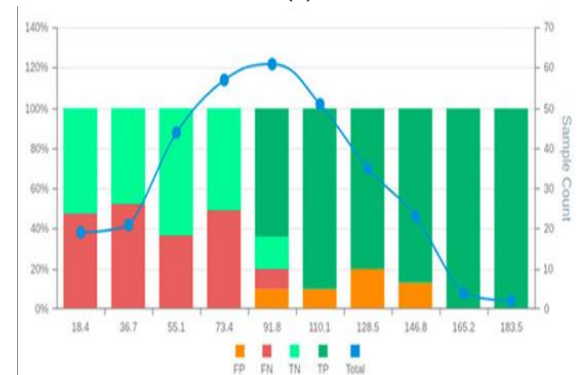
The figure legend shows the distributions of false positives (FP), false negatives (FN), true positives (TP), true negatives (TN), and the overall number of observations (Total) within each score interval in distinct colors. The left y-axis indicates the proportion of TP, TN, FP, and FN cases within each score category, while the right y-axis displays the number of samples included in each bin. The values labeled within each bin correspond to the upper boundary of the respective score interval.

Evaluation of the renal cancer model (RCM)

Assessment of the renal cancer candidate models demonstrated that RCM3 achieved the highest area under the curve (AUC) of 77.9%, while RCM6 achieved the highest overall accuracy of 74.3%. Despite these favorable performance metrics, RCM3 and RCM6 ranked second and sixth, respectively, in MSDscore, with scores of 45.8 and 5.2, respectively. RCM7 produced the highest MSDscore (65.2), with 67% sensitivity and 81% specificity. By comparison, RCM3 achieved 70% sensitivity and 82% specificity, whereas RCM6 demonstrated 81% sensitivity and 62% specificity. The MSDA visualization results provided additional evidence supporting the selection of RCM3 and RCM7 over RCM6, although RCM6 showed the highest accuracy. Both RCM3 and RCM7 showed no false positives in their highest two-score intervals, indicating that scores exceeding 145.0 for RCM3 and 146.8 for RCM7 could eliminate false positives (**Figure 6**). In contrast, RCM6 showed a broader spread of false positives across the scoring range, with a higher proportion at higher score values. RCM7's overall false-positive pattern was more stable and lower than RCM3's, which contributed to its superior MSDscore performance (**Figure 6**). However, RCM3 demonstrated a marginally lower false-negative rate, making it potentially preferable when minimizing false negatives is prioritized. Most score intervals contained sufficient observations for reliable interpretation, although the final two bins had limited case numbers.



(b)



(c)

Figure 6. Multiple score distribution (MSD) plots illustrating the performance of the customized renal cancer prognostic models: RCM3 (a), RCM6 (b), and RCM7 (c).

Each score interval displays the distribution of false-positive (FP), false-negative (FN), true-positive (TP), true-negative (TN), and total sample numbers (Total), with each category represented by a distinct color as indicated in the figure legend. The left y-axis shows the percentage contribution of TP, TN, FP, and FN cases within each score bin, while the right y-axis indicates the number of observations assigned to each bin. The numerical labels displayed within the bins correspond to the upper limit of each score interval.

The MSDA framework introduced in this study provides a new strategy for evaluating and selecting predictive models by addressing several limitations of conventional performance measures. By integrating the MSDscore metric with an accompanying visual assessment method, MSDA extends traditional evaluation approaches and offers a more detailed understanding of how models perform across different prediction score ranges.

MSDA framework and its practical application

Conventional metrics, including AUC, accuracy, sensitivity, and specificity, remain the primary tools for comparing machine learning models in biomedical research [12]. Nevertheless, choosing models solely on these aggregated indicators does not always guarantee performance preservation after implementation in real clinical environments [4, 22]. This limitation is especially significant in cancer prognostic applications, where the consequences of false-positive and false-negative predictions may differ substantially. Our analysis indicates that models with superior AUC or accuracy can still show undesirable characteristics in specific regions of the score spectrum, including increased error concentrations around clinically important decision points.

Automated machine learning platforms, such as TPOT (Tree-Based Pipeline Optimization Tool), highlight this challenge [23, 24]. Although these approaches efficiently explore extensive model and parameter spaces, they typically optimize a single predefined objective, such as AUC or accuracy. Consequently, models may achieve high statistical performance due to factors such as dataset imbalance exploitation or overfitting, without necessarily providing stable, interpretable, or clinically meaningful predictions [8, 23, 24]. The present study shows that combining MSDscore calculations with score-based visual evaluation adds another layer of model assessment. This approach enables the identification of characteristics such as regions with minimal classification errors, reliable high-confidence prediction areas, and imbalances in model behavior around specific thresholds, which may better align with clinical requirements.

The MSDA approach is intended to complement rather than replace established global performance metrics. It follows a sequential evaluation strategy in which models are initially screened using conventional performance criteria and subsequently examined using MSDA to investigate their error patterns across prediction score distributions. This combined process allows the selection of models that satisfy both quantitative performance requirements and practical considerations for real-world use.

Application of MSDA to transcriptomic-based breast, lung, and renal cancer prognostic models revealed meaningful differences between models with comparable AUC values. In breast cancer, BCM9 demonstrated strong overall predictive ability and a high-confidence prediction region, with no false-positive cases. In

contrast, the lung cancer model with the highest AUC, LCM5, displayed increased error accumulation at higher score values. In contrast, alternative models such as LCM3 and LCM9 showed prediction characteristics that may be more suitable for applications requiring greater specificity. Similarly, RCM7 achieved the highest MSD score among renal cancer models despite a lower AUC, reflecting a more balanced and clinically favorable distribution of prediction errors. These examples illustrate that MSDA can identify subtle performance differences that may remain undetected when relying exclusively on conventional summary metrics.

The relevance of MSDA is further emphasized by the challenges associated with transcriptomics-driven modeling. Transcriptomic datasets are characterized by high dimensionality, biological heterogeneity, noise, and relatively small sample sizes compared with the number of measured features [19, 25]. As a result, many cancer prognosis models demonstrate only moderate predictive performance, with reported AUC values commonly falling between 0.60 and 0.80 [13, 14]. Consistent with these observations, the models evaluated in this study achieved AUC values ranging from 62% to 83%, despite using AI-assisted optimization approaches. However, MSDA analysis demonstrated that models within this performance range could still contain highly reliable prediction regions, including areas with consistent classification behavior and predictable error distributions. Therefore, even models with moderate overall metrics may provide clinically useful information when evaluated according to their local score-space characteristics. MSDA improves model interpretability and robustness by revealing models that may have limited average performance but demonstrate valuable behavior under specific clinical conditions.

Benefits, limitations, and future directions of MSDA

One of the primary strengths of MSDA is its ability to display prediction errors across score intervals, allowing direct interpretation of local model behavior. This feature is particularly valuable in clinical applications because the impact of false-positive and false-negative outcomes is rarely equivalent. A model with average overall performance may still be suitable for deployment if it contains dependable prediction regions, such as score thresholds where false-positive classifications are extremely uncommon. Additionally, MSDscore summarises the spatial distribution of classification errors rather than simply measuring their frequency,

providing information regarding where errors occur within the prediction landscape. Such characteristics may be especially beneficial for clinical decision-support systems and prioritization tools where confidence and interpretability are essential.

The characteristics of MSDA also support its potential incorporation into automated machine learning workflows. Existing AutoML pipelines may occasionally favor models with artificially inflated global metrics because of narrow optimization objectives. Introducing MSDA as an additional assessment stage could improve model selection by identifying unstable, asymmetric, or poorly distributed error patterns that may not be apparent from traditional evaluation scores alone.

Despite these advantages, several limitations warrant acknowledgment. MSDA applies only to classification models that generate continuous prediction scores or probabilities and cannot be used with models that produce only binary outputs. Furthermore, dividing prediction scores into bins introduces a methodological consideration, as different bin sizes may influence the visual interpretation of score distributions. Accurate interpretation of MSD plots may also require some level of technical expertise, potentially limiting accessibility for inexperienced users. In addition, the current validation of MSDA has focused exclusively on transcriptomic datasets; further studies are required to determine its applicability to other data types, including proteomic, metabolomic, clinical, and imaging-based datasets. Nevertheless, the principal advantage of MSDA is its ability to expose local regions of model performance and provide insight into practical model reliability beyond what global performance indicators can provide.

Although developed for cancer prognostic modeling, MSDA is not restricted to this application and may be extended to any classification problem involving probabilistic prediction scores and threshold-based decisions. Potential applications include cardiovascular risk assessment, infectious disease prioritization, financial fraud detection, and industrial failure prediction. In these areas, where classification errors may have unequal consequences and interpretable confidence regions are valuable, MSDA may provide an effective supplementary evaluation strategy.

Fundamentally, MSDA promotes a transition from selecting models based solely on overall ranking toward evaluating whether models are appropriate for their intended purpose. The framework allows researchers and

practitioners to identify models that may not achieve the highest global metric values but provide more reliable and contextually appropriate predictions. By combining quantitative analysis of error distributions with visual interpretation of prediction behavior, MSDA supports more transparent, robust, and application-focused model selection.

Future investigations should include formal sensitivity analyses and systematic comparisons between MSDscore and established composite evaluation measures, including the Matthews Correlation Coefficient and Brier score [26, 27]. Such studies would provide additional evidence regarding MSDscore performance, comparative value, and suitability across different classification problems. Further validation using diverse datasets, machine learning algorithms, and modeling frameworks will also be important to determine the broader generalisability of the MSDA methodology. In addition, developing software implementations with user-adjustable parameters, including score bin numbers and weighting schemes, would facilitate wider adoption across academic research and industrial data science applications.

Conclusion

This study introduces a new framework for assessing and selecting score-based predictive models by integrating a composite performance metric (MSDscore) with detailed visual analysis of prediction score distributions through MSDA plots. Evaluation of 27 models across three cancer types showed that this approach can reveal important differences between models that conventional global metrics such as AUC and accuracy miss. By examining model behavior across local score regions, MSDA provides additional information regarding prediction reliability, error distribution, and the circumstances under which misclassifications occur.

The proposed methodology is designed as a complementary tool rather than a replacement for established evaluation metrics. Instead of selecting a single optimal model based solely on maximum global performance values, MSDA enables a final assessment of type I and type II error patterns across candidate models' prediction score ranges. Its utility was demonstrated in cancer prognosis using transcriptomic datasets, where the framework identified clinically relevant differences in model behavior beyond standard statistical comparisons.

Although the present study focused on cancer prognostic applications, the proposed approach may apply to other binary classification models that rely on continuous scoring systems and threshold-based decision rules. Incorporating this composite evaluation strategy into existing model development pipelines may improve the robustness, interpretability, and real-world suitability of predictive models across biomedical and other applied domains.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Strzelecki M, Badura P. Machine learning for biomedical application. *Appl Sci*. 2022;12:2022. doi:10.3390/app12042022
2. Kourou K, Exarchos KP, Papaloukas C, Sakaloglou P, Exarchos T, Fotiadis DI. Applied machine learning in cancer research: a systematic review for patient diagnosis, classification and prognosis. *Comput Struct Biotechnol J*. 2021;19:5546–55. doi:10.1016/j.csbj.2021.10.006
3. Pais RJ. Predictive modelling in clinical bioinformatics: key concepts for startups. *BioTech*. 2022;11:35. doi:10.3390/biotech11030035
4. Mann M, Kumar C, Zeng WF, Strauss MT. Artificial intelligence for proteomics and biomarker discovery. *Cell Syst*. 2021;12:759–70. doi:10.1016/j.cels.2021.06.006
5. Battineni G, Sagaro GG, Chinatalapudi N, Amenta F. Applications of machine learning predictive models in the chronic disease diagnosis. *J Pers Med*. 2020;10:21. doi:10.3390/jpm10020021
6. Assel M, Vickers A. Biomarker evaluation and clinical development. *Soc Int Urol J*. 2020;1:16–22. doi:10.48083/XIQR6849
7. Gopalakrishna G, Langendam M, Scholten R, Bossuyt P, Leeflang M, Noel-Storr A, et al. Methods for evaluating medical tests and biomarkers. *Diagn Progn Res*. 2017;1(Suppl 1):7. doi:10.1186/s41512-017-0007-6
8. Telikani A, Gandomi AH, Tahmassebi A, Banzhaf W. Evolutionary machine learning: a survey. *ACM Comput Surv*. 2021;54:1–35. doi:10.1145/3467477
9. Kim H, Kwon HJ, Kim ES, Kwon S, Suh KJ, Kim SH, et al. Comparison of the predictive power of a combination versus individual biomarker testing in non-small cell lung cancer patients treated with immune checkpoint inhibitors. *Cancer Res Treat*. 2022;54:424–33. doi:10.4143/crt.2021.505
10. Boeri C, Chiappa C, Galli F, De Berardinis V, Bardelli L, Carcano G, et al. Machine learning techniques in breast cancer prognosis prediction: a primary evaluation. *Cancer Med*. 2020;9:3234–43. doi:10.1002/cam4.2811
11. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol*. 2010;5:1315–6. doi:10.1097/JTO.0b013e3181ec173d
12. Dankers FJWM, Traverso A, Wee L, van Kuijk SMJ. Prediction modeling methodology. In: *Fundamentals of Clinical Data Science*. Cham: Springer; 2019. p. 101–20. doi:10.1007/978-3-319-99713-1_8
13. Pais RJ, Lopes F, Parreira I, Silva M, Silva M, Moutinho MG. Predicting cancer prognostics from tumour transcriptomics using an auto machine learning approach. *Med Sci Forum*. 2023;22:6. doi:10.3390/msf2023022006
14. Yang D, Ma X, Song P. A prognostic model of non small cell lung cancer based on TCGA and ImmPort databases. *Sci Rep*. 2022;12:437. doi:10.1038/s41598-021-04487-6
15. Zhang J, Wang Y, Molino P, Li L, Ebert DS. Manifold: a model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE Trans Vis Comput Graph*. 2018;25:364–73. doi:10.1109/TVCG.2018.2864499
16. Wang Q, Alexander W, Pegg J, Qu H, Chen M. HypoML: visual analysis for hypothesis-based evaluation of machine learning models. *IEEE Trans Vis Comput Graph*. 2020;27:1417–26. doi:10.1109/TVCG.2020.3030088
17. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak*. 2006;26:565–74. doi:10.1177/0272989X06295361
18. Filho UL, Pais TA, Pais RJ. Facilitating “omics” for phenotype classification using a user-friendly AI-driven platform: application in cancer

- prognostics. *BioMedInformatics*. 2023;3:1071–82. doi:10.3390/biomedinformatics3040055
19. Edwards NJ, Oberti M, Thangudu RR, Cai S, McGarvey PB, Jacob S, et al. The CPTAC data portal: a resource for cancer proteomics research. *J Proteome Res*. 2015;14:2707–13. doi:10.1021/pr501254j
20. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhor G, et al. A pathology atlas of the human cancer transcriptome. *Science*. 2017;357:2507. doi:10.1126/science.aan2507
21. Pais RJ. Simulation of multiple microenvironments shows a pivot role of RPTs on the control of epithelial-to-mesenchymal transition. *Biosystems*. 2020;198:104268. doi:10.1016/j.biosystems.2020.104268
22. Swan AL, Mobasher A, Allaway D, Liddell S, Bacardit J. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *OMICS*. 2013;17:595–610. doi:10.1089/omi.2013.0017
23. Le TT, Fu W, Moore JH. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics*. 2020;36:250–6. doi:10.1093/bioinformatics/btz470
24. Olson RS, Urbanowicz RJ, Andrews PC, Lavender NA, Kidd LC, Moore JH. Automating biomedical data science through tree-based pipeline optimization. In: *Applications of Evolutionary Computation. Lecture Notes in Computer Science*. Vol 9597. Cham: Springer; 2016. p. 123–37. doi:10.1007/978-3-319-31204-0_9
25. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. *Science*. 2015;347:1260419. doi:10.1126/science.1260419
26. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res*. 2019;3:18. doi:10.1186/s41512-019-0064-0
27. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21:6. doi:10.1186/s12864-019-6413-7