Hyeon Seok Seok, Yuna Choi, Shinae Yu, Kyung-Hwa Shin, Sollip Kim* and Hangsik Shin*

Machine learning-based delta check method for detecting misidentification errors in tumor marker tests

https://doi.org/10.1515/cclm-2023-1185
Received October 23, 2023; accepted November 30, 2023;
published online December 14, 2023

Abstract

Objectives: Misidentification errors in tumor marker tests can lead to serious diagnostic and treatment errors. This study aims to develop a method for detecting these errors using a machine learning (ML)-based delta check approach, overcoming limitations of conventional methods.

Methods: We analyzed five tumor marker test results: alpha-fetoprotein (AFP), cancer antigen 19-9 (CA19-9), cancer antigen 125 (CA125), carcinoembryonic antigen (CEA), and prostate-specific antigen (PSA). A total of 246,261 records were used in the analysis. Of these, 179,929 records were used for model training and 66,332 records for performance evaluation. We developed a misidentification error detection model based on the random forest (RF) and deep neural network (DNN) methods. We performed an *in silico* simulation with 1 % random sample shuffling. The performance of the developed models was evaluated and compared to conventional delta check methods such as delta percent change (DPC), absolute DPC (absDPC), and reference change values (RCV).

Results: The DNN model outperformed the RF, DPC, absDPC, and RCV methods in detecting sample misidentification errors. It achieved balanced accuracies of 0.828, 0.842, 0.792, 0.818, and 0.833 for AFP, CA19-9, CA125, CEA, and PSA, respectively. Although the RF method performed better than DPC and absDPC, it showed similar or lower performance compared to RCV.

Conclusions: Our research results demonstrate that an ML-based delta check method can more effectively detect sample misidentification errors compared to conventional delta check methods. In particular, the DNN model demonstrated superior and stable detection performance compared to the RF, DPC, absDPC, and RCV methods.

Keywords: artificial intelligence; autoverification; deep neural network; delta check; machine learning; tumor markers

Introduction

Tumor markers are helpful in diagnosing, prognosticating, and monitoring cancer treatment [1]. Accurate tumor marker test results are essential for proper use. We analyzed five commonly used tumor markers: alpha-fetoprotein (AFP), carbohydrate antigen 19-9 (CA 19-9), cancer antigen 125 (CA-125), carcinoembryonic antigen (CEA), and prostate-specific antigen (PSA). AFP is used to aid in diagnosis of hepatocellular carcinoma and to screen high-risk patients, and low AFP values after treatment are associated with a favorable prognosis. CA 19-9 and CA-125 are used to aid in diagnosis of pancreatic ductal adenocarcinoma and monitor treatment efficacy for pancreatic and ovarian cancer, respectively. CEA is mostly used for monitoring the treatment efficacy of metastatic colorectal cancer and is used as an adjunct to diagnosis due to its high specificity for colorectal cancer, and a high CEA measurement is a marker of poor prognosis. PSA is used to aid in the diagnosis and monitoring of the treatment efficacy of prostate cancer.

Sollip Kim and Hangsik Shin contributed equally to this work.

***Corresponding authors: Sollip Kim**, MD, PhD, Department of Laboratory Medicine, Asan Medical Center, University of Ulsan College of Medicine, 88, Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea, Phone: 82 2 3010 4553, Fax: +82 2 2045 3081, E-mail: sollip_kim@amc.seoul.kr, Web of Science ResearcherID: E-8546-2011. https://orcid.org/0000-0003-0474-5897; and **Hangsik Shin**, PhD, Department of Digital Medicine, Asan Medical Center, University of Ulsan College of Medicine, 88, Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea, Phone: 82 2 3010 2099, Fax: +82 2 3010 4182, E-mail: hangsik.shin@amc.seoul.kr, Web of Science ResearcherID: JFL-1492-2023. https://orcid.org/0000-0002-3353-0310

Hyeon Seok Seok, Interdisciplinary Program of Biomedical Engineering, Chonnam National University, Yeosu, Republic of Korea. https://orcid.org/0000-0001-8144-2190

Yuna Choi, Department of Laboratory Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea. https://orcid.org/0000-0001-7869-5015

Shinae Yu, Department of Laboratory Medicine, Haeundae Paik Hospital, Inje University College of Medicine, Busan, Republic of Korea. https://orcid.org/0000-0002-9527-5853

Kyung-Hwa Shin, Department of Laboratory Medicine and Biomedical Research Institute, Pusan National University Hospital, Busan, Republic of Korea. https://orcid.org/0000-0002-8454-4448

In the clinical laboratory testing process, preanalytical errors can occur during sample collection, handling, or transportation, leading to erroneous results [2]. These errors account for 60–70 % of all testing errors [3], with the majority (about 75 %) attributed to sample quality issues such as hemolysis or sample clotting [2]. These errors are mostly detected during the testing process and the clinical laboratory either rejects the sample based on the degree of hemolysis or proceeds to the next process with a comment on the sample quality (e.g., "hemolysis may affect the test results") in the test report, along with the test results [4]. However, sample misidentification errors – although comprising only about 0.3 % of all preanalytical errors [2] – can lead to serious consequences when one patient's results are erroneously associated with another's, leading to diagnostic and therapeutic mistakes.

To avoid preanalytical errors, clinical laboratory investigators use various verification methods before reporting test results. One strategy is a delta check, which alerts laboratory personnel to potential errors if the difference between the last and current test results exceeds a certain threshold. The conventional delta check method identifies results that deviate from statistically established extreme values stemming from either patients' biological variations or variations in laboratory instruments. The conventional delta check method is considered effective in detecting preanalytical errors such as sample misidentification, sample contamination, and hemolysis [5]; however, the sensitivity of the conventional delta check method remains around 20 % [6]. Likewise, in our previous study, when practical delta check limits were established for five tumor markers by the conventional, delta percent change (DPC), and absolute DPC (absDPC) methods, the sensitivity was 20–50 % depending on the clinical setting and test items [7].

To address this issue, research is underway on a machine learning (ML)-based delta check method that has demonstrated superior performance compared to traditional methods [8, 9]. ML can learn complex relationships and patterns through non-linear learning mechanisms, and this approach can detect subtle and adaptive changes that are difficult to detect with conventional statistical approaches [10]. Unlike traditional methods, ML techniques can learn from data distributions and patterns to build optimized models, which help to improve error detection. While a few ML-based delta check studies have been reported for general chemistry and general hematology tests [8, 11, 12], none have been reported for tumor markers. Therefore, this study aims to develop and validate an ML-based delta check method for detecting sample misidentification errors in clinical laboratories. We developed a random forest (RF) model and a deep learning model to detect sample misidentification errors using retrospectively collected test results of five tumor markers, AFP, CA19-9, CA125, CEA, and PSA, and compared their performance with conventional delta check methods such as DPC, absDPC, and reference change values (RCV).

Materials and methods

Data collection

Data used in this study are the same as those used in our previous research [7]. Five tumor markers (AFP, CA19-9, CA125, CEA, and PSA) were retrospectively collected from a laboratory information management system (LIMS) that stored the results of tests performed using a Roche Cobas C-8000 (Roche Diagnostics GmbH, Mannheim, Germany) at the Pusan National University Hospital (Busan, Republic of Korea), Haeundae Paik Hospital (Busan, Republic of Korea), and Ilsan Paik Hospital (Goyang, Republic of Korea) from Jan. 2020 to Dec. 2021.

Data collected include current result (the current measurement result), previous result (the result of the last test performed within the last two years), age, sex, test requesting department, date reported, and patient class from LIMS, excluding any missing data. In our study, we utilized previous and current results as input data for model development. However, variables such as age, sex, test requesting department, and date reported were omitted from the input data. This exclusion was deliberate, as these factors were deemed unrelated to sample misidentification errors or were considered insufficient in capturing real-world environmental influences. Patient classes are used for subgroup analysis. Patient class consisted of patients who underwent health screening (H), outpatients (O), and emergency patients or inpatients (I). Data with test results exceeding the analytic measuring interval (AMI) of the analysis system was excluded. The applied AMIs were 0.908–1,210 μg/L for AFP, 0.6–5,000 KU/L for CA125, 0.6–1,000 KU/L for CA19-9, 0.2–1,000 μg/L for CEA, and 0.003–100 μg/L for PSA [7]. Table 1 shows the dataset information for each of the five tumor markers used in this study. The development set (D-set) consists of 179,929 records for the first 18 months (Jan. 2020-Jun. 2021) to develop the delta check method and the test set (T-set) consists of 66,332 records for the last 6 months (Jul. 2021-Dec. 2021) to evaluate the delta check method.

The data used in this study was collected after approval by the Ethics Review Boards of each institution, and the requirement for human consent was exempted as a retrospective study (PNUH 2210-023-120, HPIRB 2022-09-017, ISPAIK 2022-09-031).

Overall process of sample misidentification error detection

Figure 1 shows the overall model development and validation processes. We used random forest (RF) and deep neural network (DNN), which are frequently used in feature-based binary classification, as ML models for sample misidentification error detection. RF is a tree-based classification and regression model that performs well in feature-based analysis [13]. DNN is an artificial neural network that contains multiple hidden layers and is excellent at learning nonlinear relationships [14]. The performance of the ML-based sample misidentification error detection

Table 1: Dataset description.

	Number of samples	
	Development set	**Test set**
Data collection period		
	Jan. 2020–Jun. 2021	Jul. 2021–Dec. 2021
Tumor marker		
AFP	44,255	16,229
CA125	22,735	8,187
CA19-9	39,079	14,398
CEA	50,274	18,641
PSA	23,586	8,877
Patient class		
H	49,852	17,955
O	115,844	43,070
I	14,233	5,307
Total	179,929	66,332

AFP, alpha-fetoprotein; CA 19-9, carbohydrate antigen 19-9; CA-125, cancer antigen 125; CEA, carcinoembryonic antigen; PSA, prostate-specific antigen; H, health screening; O, outpatients; I, emergency patients or inpatients.

method was compared with the conventional verification method using DPC, absDPC, and RCV. All models were developed using the D-set and tested using the T-set. ML models were trained using data that simulated "sample shuffling" by randomly shuffling a certain percentage of the current results, while the delta check limits of conventional methods were derived using data without sample shuffling. ML model's performance was evaluated without considering patient class. The conventional model's performance was performed in two directions. The first was to derive delta check limits from all D-sets and apply them to all T-sets without considering patient class, and the second was to consider patient class and derive delta check limits from patient classes H, O, and I in the D-sets and evaluate their performance independently for the corresponding patient classes in the T-sets. The comparative study between the proposed ML-based model and the conventional model is shown in Figure 2. Python 3.9 (Python Software Foundation) was used for model implementation, preprocessing, and DPC, absDPC, and RCV calculations, and TensorFlow (ver. 2.14.0), Keras (ver. 2.13.1), and scikit-learn (ver. 1.2.2) were additionally used for RF and DNN model implementations.

ML-based delta check model

To simulate sample misidentification errors, we randomly shuffled 1 % of the current results in the D-set. We then labeled these shuffled results as "misidentified," denoted by 1, while the correctly identified results were labeled as "identified," denoted by 0. The input data was log transformed to reduce the impact of extreme values and normalize the data. Log transformation is a preprocessing method often used for continuous data in model training in ML. Log transformation is particularly useful for transforming skewed long-tailed distributions into something that more closely resembles a normal distribution, and is known to help train more sophisticated models by improving the normality of the input data [15, 16]. Robust scaling was applied to set the median value (Q2) of each feature to 0 and normalize it based on the interquartile range, which is the difference between the 1st (25th quantile, Q1) and 3rd quartiles (75th quantile, Q3): normalized value=(value-Q2)/(Q3-Q1).

While implementing RF and DNN models, each model's hyperparameters were optimized by random search. In this process, balanced accuracy was used as the performance metric to mitigate the impact of class imbalance. The detailed model architecture and hyperparameters are described in Supplementary Material 1. Then, we performed 5-fold cross-validation to verify the more generalized performance of the model. In 5-fold cross validation, the dataset was divided into 5 evenly sized folds, and the process of training the model on each 1-fold and evaluating it on the remaining folds was performed independently 5 times by changing the evaluation fold, and finally the average was

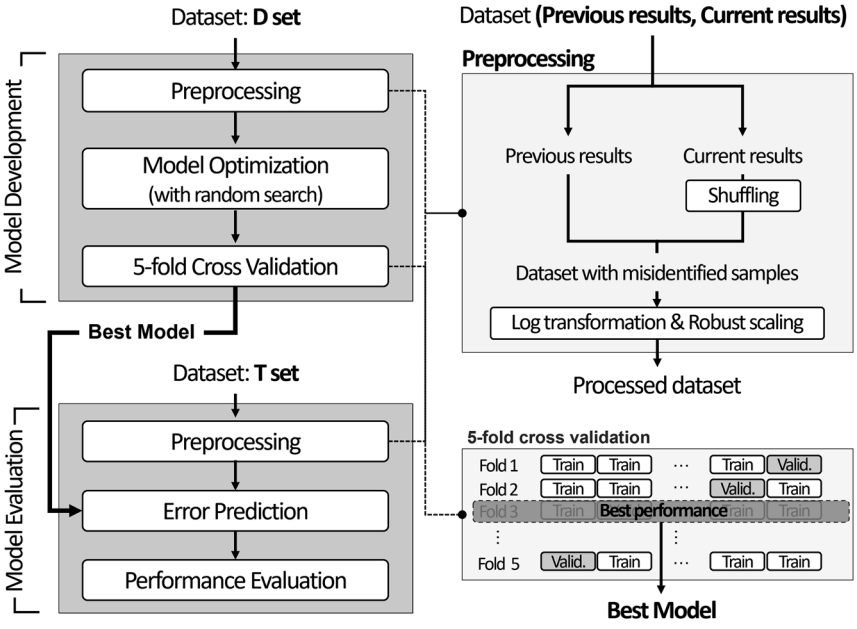

Figure 1: Overall process of model development and evaluation.

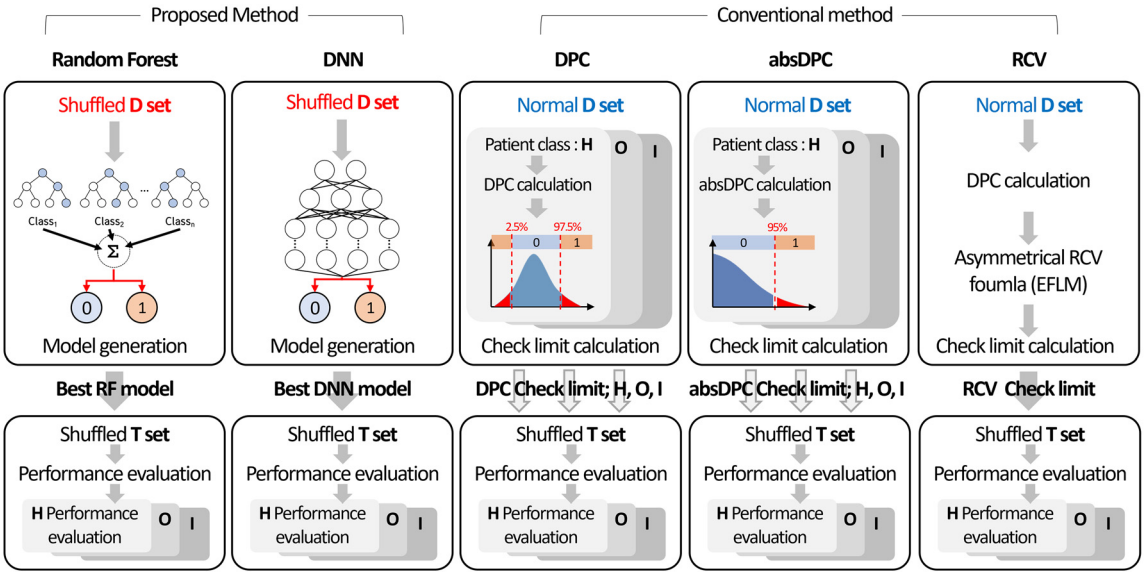

Figure 2: Comparative study between the proposed ML-based model and the conventional model; RF, random forest; DNN, deep neural network; DPC, delta percent change; absDPC, absolute delta percent change; RCV, reference change values; H, health screening; O, outpatients; I, emergency patient or inpatients.

taken as the general model's performance. The model at the fold with the highest area under the curve of the receiver operating characteristic curve (AUROC) in the evaluation was then selected as the "best model". Finally, the model's error detection performance was evaluated by applying T-set to the best model.

Conventional delta check methods

Conventional delta checks were performed using the DPC, absDPC, and RCV check limits established in our previous work [7]. DPC was calculated as the ratio of the change between the previous and current test results divided by the previous test results (Eq. (1)), absDPC was presented as the absolute value of the DPC result (Eq. (2)), and RCV was derived from the asymmetrical RCV formula based on the biological variation database of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) [17].

$$\mathrm{DPC}(\%) = \frac{X_{\mathrm{current}} - X_{\mathrm{previous}}}{X_{\mathrm{previous}}} \times 100 \quad (1)$$

$$\mathrm{absDPC}(\%) = \left| \frac{X_{\mathrm{current}} - X_{\mathrm{previous}}}{X_{\mathrm{previous}}} \right| \times 100 \quad (2)$$

The delta check limit was calculated for each H, O, and I within the D-set. DPC set the limit to 2.5 % of both extremes of the value distribution, and absDPC set the limit to 95 %.

Performance evaluation

The ML model's performance was evaluated using the permutation test, which involves repeatedly evaluating the performance of a model while randomly re-sampling to obtain the average performance. In this study, the process of training and validating the model was repeated 1,000 times, with 1 % of the data randomly shuffled in each iteration. We evaluated the sample misidentification error detection performance of DPC, absDPC, and RCV (conventional method) using a T-set that contained a 1 % sample misidentification error. Initially, we employed delta check limits derived from the entire D-set to assess the entire T-set. Additionally, we conducted a separate analysis for patient classes H, I, and O within the T-set, utilizing delta test limits derived from the D-set specific to each patient class.

For the performance evaluation, we utilized the AUROC, balanced accuracy, sensitivity, and specificity. AUROC serves as a critical indicator of a model's classification capability, with values ranging from 0 to 1, where higher values indicate superior performance. Balanced accuracy is calculated as the average of sensitivity and specificity and provides a more accurate assessment of a model's performance, particularly in datasets with imbalanced class distributions; it signifies how well the model balances its classification of different classes. Sensitivity measures the model's ability to accurately detect true positives, while specificity assesses its capacity to correctly identify true negatives.

Results

Performance evaluation of the delta check method

The sample misidentification error detection performance of the developed models is shown in Figure 3. In Figure 3, the blue and red solid lines represent the average ROC curves, and the light blue and light red areas represent the standard deviation ranges of RF and DNN, respectively. The performance of DPC, absDPC, and RCV methods is denoted by each symbol, and the coordinates mean (sensitivity, 1-specificity). The classification performance of the RF model was 0.791,

0.816, 0.751, 0.797, and 0.787 AUROC, lower than that of the DNN, which was 0.831, 0.877, 0.829, 0.838, and 0.860 AUROC for AFP, CA19-9, CA125, CEA, and PSA, respectively. Comparisons using AUROC showed that DNN generally performed better for all tumor markers. RF showed better detection performance than the DPC and absDPC methods but lower performance than the DNN and RCV methods.

Table 2 shows the sample misidentification error detection performance of RF, DNN, DPC, absDPC, and RCV methods by tumor marker type. RF detected the sample misidentification error with balanced accuracy of 0.753, 0.721, 0.746, 0.771, and 0.575 for AFP, CA19-9, CA125, CEA, and PSA, respectively. However, RF performed worse in CA19-9 and PSA compared to RCV, with the lowest performance (0.575 of balanced accuracy) in PSA. Meanwhile, the sample misidentification error detection accuracy of the DNN method for AFP, CA19-9, CA125, CEA, and PSA tumor markers were 0.828, 0.842, 0.792, 0.818, and 0.833, respectively, which were the highest of any other methods; moreover, the sensitivity was 0.909, 0.852, 0.832, 0.804, and 0.808, better than other methods except for PSA.

Performance evaluation of the delta check method by patient class

Figure 4 shows the performance evaluation results of each model by patient class. Rows correspond to tumor markers and are AFP, CA19-9, CA125, CEA, and PSA from top to bottom, respectively; columns correspond to patient class and are H, O, and I from left to right, respectively; the blue solid line represents the ROC curve of the RF, while the red solid line represents the ROC curve of the DNN. The evaluation results of the DPC, absDPC, and RCV methods are indicated by the coordinates corresponding to (sensitivity, 1-specificity). The DNN performed best overall for all tumor markers and all patient classes; however, in patient class I, it performed similar to or slightly lower than the RCV. The RF performed

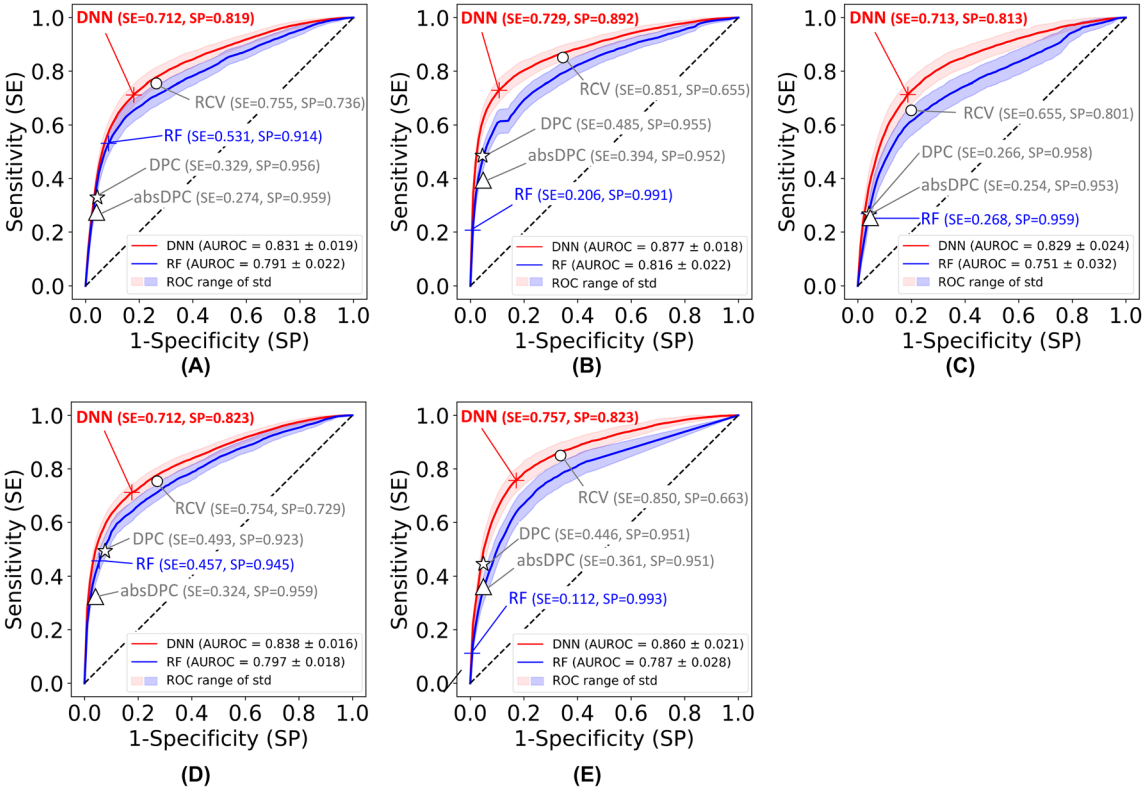

Figure 3: ROC curves of the RF and DNN models and conventional methods: (A) AFP, (B) CA19-9, (C) CA125, (D) CEA, (E) PSA. The maker means the best threshold of RF, DNN, DPC, absDPC, and RCV. AFP, alpha-fetoprotein; CA 19-9, carbohydrate antigen 19-9; CA-125, cancer antigen 125; CEA, carcinoembryonic antigen; PSA, prostate-specific antigen; RF, random forest; DNN, deep neural network; DPC, delta percent change; absDPC, absolute delta percent change; RCV, reference change values; SE, sensitivity; SP, specificity; ROC, receiver operating characteristic.

Table 2: Performance comparison for each delta check model in detecting sample misidentification errors for tumor markers. Numbers represent the mean (95 % CI).

Tumor marker	Method	Balanced accuracy	Sensitivity	Specificity
AFP	RF	0.753 (0.742–0.764)	0.616 (0.594–0.638)	0.889 (0.889–0.889)
	DNN	0.828 (0.821–0.834)	0.909 (0.897–0.922)	0.746 (0.745–0.746)
	DPC	0.643 (0.639–0.647)	0.329 (0.321–0.338)	0.956 (0.956–0.956)
	absDPC	0.616 (0.613–0.620)	0.274 (0.267–0.281)	0.959 (0.959–0.959)
	RCV	0.745 (0.741–0.749)	0.755 (0.747–0.763)	0.736 (0.736–0.736)
CA19-9	RF	0.721 (0.711–0.731)	0.482 (0.463–0.502)	0.960 (0.960–0.960)
	DNN	0.842 (0.834–0.849)	0.852 (0.838–0.867)	0.831 (0.831–0.831)
	DPC	0.720 (0.716–0.724)	0.485 (0.477–0.494)	0.955 (0.955–0.955)
	absDPC	0.673 (0.669–0.677)	0.394 (0.387–0.401)	0.952 (0.952–0.952)
	RCV	0.753 (0.750–0.756)	0.851 (0.845–0.857)	0.655 (0.655–0.655)
CA125	RF	0.746 (0.732–0.760)	0.610 (0.582–0.637)	0.883 (0.882–0.883)
	DNN	0.792 (0.780–0.805)	0.832 (0.807–0.858)	0.753 (0.752–0.753)
	DPC	0.612 (0.606–0.618)	0.266 (0.254–0.278)	0.958 (0.958–0.958)
	absDPC	0.603 (0.599–0.608)	0.254 (0.246–0.263)	0.953 (0.953–0.953)
	RCV	0.728 (0.723–0.734)	0.655 (0.644–0.666)	0.801 (0.801–0.802)
CEA	RF	0.771 (0.764–0.779)	0.649 (0.635–0.664)	0.893 (0.893–0.893)
	DNN	0.818 (0.812–0.824)	0.804 (0.792–0.816)	0.832 (0.832–0.833)
	DPC	0.708 (0.704–0.712)	0.493 (0.485–0.501)	0.923 (0.923–0.923)
	absDPC	0.641 (0.637–0.645)	0.324 (0.316–0.331)	0.959 (0.959–0.959)
	RCV	0.741 (0.738–0.745)	0.754 (0.748–0.761)	0.729 (0.729–0.729)
PSA	RF	0.575 (0.568–0.581)	0.165 (0.152–0.177)	0.985 (0.985–0.985)
	DNN	0.833 (0.827–0.840)	0.808 (0.794–0.821)	0.859 (0.859–0.859)
	DPC	0.699 (0.693–0.704)	0.446 (0.435–0.458)	0.951 (0.951–0.951)
	absDPC	0.655 (0.651–0.658)	0.361 (0.353–0.368)	0.949 (0.948–0.949)
	RCV	0.757 (0.753–0.761)	0.850 (0.842–0.858)	0.663 (0.663–0.663)

CI, confidence interval; AFP, alpha-fetoprotein; CA 19-9, carbohydrate antigen 19-9; CA-125, cancer antigen 125; CEA, carcinoembryonic antigen; PSA, prostate-specific antigen; RF, random forest; DNN, deep neural network; DPC, delta percent change; absDPC, absolute delta percent change; RCV, reference change values.

worse than the DNN and RCV in all cases and was similar to the DPC and absDPC. The detailed performance metrics of RF, DNN, DPC, absDPC, and RCV according to patient class are shown in Tables 3–5.

Discussion

In this study, we developed delta check methods based on RF and DNN models for each of the five tumor markers. We evaluated their performance in detecting sample misidentification errors and compared them with conventional methods like DPC, absDPC, and RCV. This is the first study to develop an ML-based delta check method for tumor marker tests. Additionally, we developed the ML method using the same raw data as in [7], enabling a fair comparison with the conventional delta check method. Sensitivity and specificity are crucial metrics for evaluating model performance. However, optimizing both simultaneously is challenging due to their trade-off relationship. Thus, criteria should be set based on the situation. From delta check perspective, sensitivity refers to detecting an abnormal sample when a sample misidentification error has occurred. However, higher sensitivity can increase false positives, leading to unnecessary retests or system overhauls and increasing laboratory workload. Meanwhile, specificity refers to considering an error-free sample as normal. However, higher specificity can lead to more false negatives. Therefore, it is crucial to ensure that both metrics perform at an acceptable level for clinical laboratories rather than biasing a model towards either sensitivity or specificity. Notably, the ROC curve of DNN outperforms that of conventional methods (see Figure 3). Giving that sensitivity and specificity depend on cut-off values but are ultimately determined by the ROC curve, DNN's performance is considered superior to conventional methods despite trade-off. However, the target values for sensitivity and specificity may vary across laboratories, so optimization may require adjusting cut-off values based on laboratory policy.

The superior performance of the DNN model over the RF model in detecting sample misidentification errors can be attributed to differences in the underlying working principles of these two models. DNN iteratively performs weighted

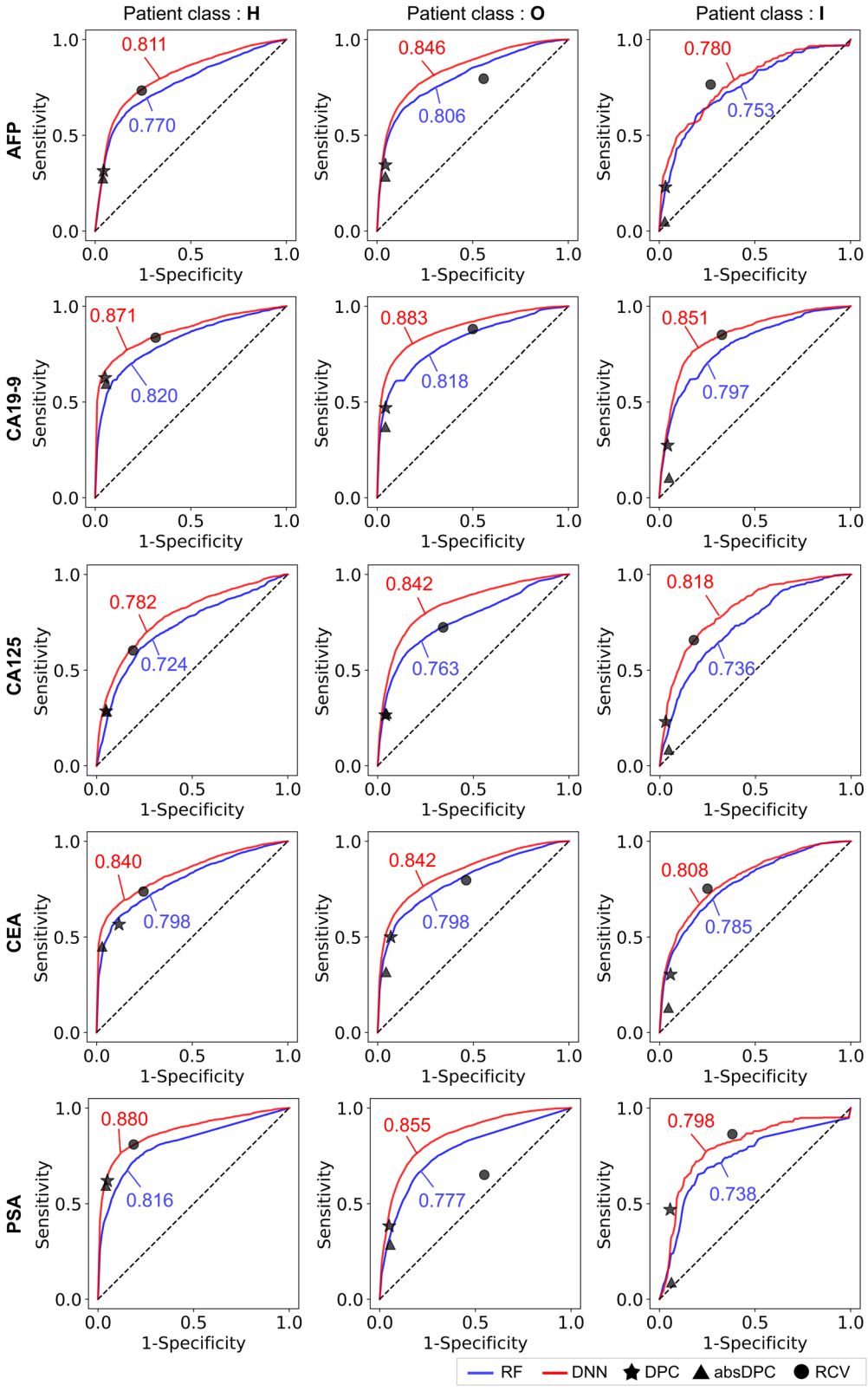

Figure 4: Performance comparison of ML-based methods and conventional methods according to the patient class. Solid lines represent the ROC curve and numbers represent the AUROC; rows represent the type of tumor marker; column represent the patient class. AFP, alpha-fetoprotein; CA 19-9, carbohydrate antigen 19-9; CA-125, cancer antigen 125; CEA, carcinoembryonic antigen; PSA, prostate-specific antigen; RF, random forest; DNN, deep neural network; DPC, delta percent change; absDPC, absolute delta percent change; RCV, reference change values; H, health screening; O, outpatients; I, emergency patients or inpatients; ML, machine learning; ROC, receiver operating characteristic; AUROC, area under the receiver operating characteristic.

Table 3: Performance comparison for each delta check model in detecting sample misidentification errors for tumor markers in patient class H (health screening). Delta check methods based on RF and DNN were developed using a 1 % randomly shuffled total D-set, without regard to patient class. Conventional delta check limits for DPC, absDPC, and RCV were derived using the non-shuffled D-set from the patient class H. Numbers represent the mean (95 % CI).

Tumor marker	Method	Balanced accuracy	Sensitivity	Specificity
AFP	RF	0.713 (0.706–0.720)	0.519 (0.506–0.533)	0.906 (0.906–0.907)
	DNN	0.753 (0.747–0.759)	0.698 (0.687–0.710)	0.808 (0.808–0.808)
	DPC	0.635 (0.629–0.641)	0.314 (0.302–0.327)	0.956 (0.956–0.956)
	absDPC	0.618 (0.612–0.623)	0.275 (0.264–0.286)	0.960 (0.960–0.961)
	RCV	0.746 (0.741–0.752)	0.735 (0.724–0.746)	0.757 (0.757–0.757)
CA19-9	RF	0.599 (0.592–0.607)	0.205 (0.191–0.220)	0.993 (0.993–0.993)
	DNN	0.810 (0.802–0.818)	0.710 (0.694–0.726)	0.910 (0.910–0.910)
	DPC	0.788 (0.779–0.796)	0.627 (0.610–0.644)	0.948 (0.948–0.948)
	absDPC	0.769 (0.761–0.777)	0.596 (0.579–0.612)	0.943 (0.943–0.943)
	RCV	0.760 (0.754–0.766)	0.836 (0.824–0.848)	0.684 (0.684–0.684)
CA125	RF	0.607 (0.596–0.619)	0.282 (0.259–0.306)	0.932 (0.932–0.932)
	DNN	0.719 (0.707–0.732)	0.685 (0.660–0.710)	0.753 (0.753–0.753)
	DPC	0.618 (0.608–0.629)	0.286 (0.265–0.307)	0.950 (0.950–0.950)
	absDPC	0.616 (0.606–0.625)	0.285 (0.265–0.304)	0.947 (0.947–0.947)
	RCV	0.707 (0.695–0.720)	0.604 (0.579–0.629)	0.810 (0.810–0.810)
CEA	RF	0.709 (0.701–0.717)	0.458 (0.441–0.474)	0.960 (0.960–0.961)
	DNN	0.775 (0.769–0.782)	0.690 (0.677–0.704)	0.860 (0.860–0.860)
	DPC	0.725 (0.717–0.733)	0.566 (0.551–0.582)	0.883 (0.883–0.883)
	absDPC	0.711 (0.702–0.719)	0.450 (0.433–0.466)	0.971 (0.971–0.972)
	RCV	0.746 (0.739–0.753)	0.737 (0.722–0.751)	0.755 (0.755–0.755)
PSA	RF	0.566 (0.559–0.574)	0.138 (0.124–0.153)	0.995 (0.995–0.995)
	DNN	0.818 (0.809–0.827)	0.732 (0.715–0.749)	0.904 (0.904–0.904)
	DPC	0.785 (0.775–0.795)	0.619 (0.599–0.639)	0.951 (0.951–0.951)
	absDPC	0.777 (0.766–0.787)	0.595 (0.574–0.615)	0.959 (0.959–0.959)
	RCV	0.812 (0.803–0.820)	0.809 (0.792–0.826)	0.814 (0.814–0.815)

CI, confidence interval; AFP, alpha-fetoprotein; CA 19-9, carbohydrate antigen 19-9; CA-125, cancer antigen 125; CEA, carcinoembryonic antigen; PSA, prostate-specific antigen; RF, random forest; DNN, deep neural network; DPC, delta percent change; absDPC, absolute delta percent change; RCV, reference change values.

sum operations on input data to generate a continuous output value, which is subsequently evaluated for errors using decision criteria. In contrast, RF relies on the continuous segmentation of randomly selected features according to specific criterion values to predict the dependent variable. Given that delta check fundamentally involves "analyzing the difference between two inputs," it is noteworthy that, unlike DNNs, tree-based models split the tree based on absolute values without performing operations between input features. This can lead to significant performance variations if the range of input values is diverse or extensive, affecting the number or depth of the tree. Moreover, RF models can result in overfitting or unstable prediction outcomes when dealing with a small number of input features due to the random feature selection process. In our study, we utilized only two test results, the previous and current results, as inputs to the model, and it is presumed that DNN offers an advantage over the RF model in achieving robust performance under these circumstances.

DPC and absDPC establish delta check limits for detecting sample misidentification errors by using the distribution of results in a specific patient group to statistically. For non-standardized or non-harmonized tests, measurements vary depending on the laboratory method used and reagents and calibration materials manufacturer, especially for tumor markers [18]. Delta check limits should not be adopted directly from different laboratories or clinical settings. Since RCVs are set using intra-individual variability and laboratory imprecision, they must be set considering each clinical condition. We derived delta check limits of DPC, absDPC, and RCV by patient class and applied them to each patient class to evaluate the sample misidentification error detection performance. RF and DNN were evaluated by applying the trained model to all patient classes without distinguishing patient classes. The results showed that when evaluated by patient class, DNN outperformed the conventional delta test methods optimized by patient class, even though they were not optimized by patient class. However, in the case of

Table 4: Performance comparison for each delta check model in detecting sample misidentification errors for tumor markers in patient class O (outpatients). Delta check methods based on RF and DNN were developed using a 1 % randomly shuffled total D-set, without regard to patient class. Conventional delta check limits for DPC, absDPC, and RCV were derived using the non-shuffled D-set from the patient class O. Numbers represent the mean (95 % CI).

Tumor marker	Method	Balanced accuracy	Sensitivity	Specificity
AFP	RF	0.731 (0.726–0.735)	0.538 (0.529–0.548)	0.923 (0.923–0.923)
	DNN	0.777 (0.772–0.782)	0.719 (0.709–0.729)	0.835 (0.835–0.835)
	DPC	0.650 (0.645–0.655)	0.344 (0.334–0.354)	0.956 (0.956–0.956)
	absDPC	0.621 (0.616–0.625)	0.284 (0.275–0.293)	0.957 (0.957–0.957)
	RCV	0.749 (0.744–0.754)	0.766 (0.757–0.775)	0.732 (0.732–0.732)
CA19-9	RF	0.599 (0.595–0.604)	0.205 (0.196–0.214)	0.993 (0.993–0.993)
	DNN	0.814 (0.809–0.819)	0.730 (0.720–0.740)	0.898 (0.898–0.899)
	DPC	0.713 (0.708–0.719)	0.470 (0.459–0.481)	0.956 (0.956–0.957)
	absDPC	0.663 (0.658–0.668)	0.370 (0.360–0.379)	0.957 (0.957–0.957)
	RCV	0.762 (0.758–0.766)	0.851 (0.843–0.859)	0.673 (0.673–0.673)
CA125	RF	0.620 (0.616–0.624)	0.270 (0.262–0.278)	0.970 (0.970–0.970)
	DNN	0.778 (0.773–0.784)	0.718 (0.707–0.729)	0.839 (0.839–0.839)
	DPC	0.612 (0.605–0.619)	0.265 (0.251–0.280)	0.959 (0.959–0.959)
	absDPC	0.613 (0.608–0.618)	0.272 (0.261–0.282)	0.954 (0.954–0.954)
	RCV	0.740 (0.733–0.746)	0.658 (0.645–0.670)	0.822 (0.822–0.822)
CEA	RF	0.700 (0.695–0.704)	0.450 (0.440–0.459)	0.950 (0.950–0.950)
	DNN	0.771 (0.767–0.775)	0.711 (0.703–0.719)	0.832 (0.832–0.832)
	DPC	0.716 (0.711–0.721)	0.500 (0.490–0.510)	0.932 (0.932–0.932)
	absDPC	0.636 (0.632–0.641)	0.317 (0.308–0.326)	0.956 (0.956–0.956)
	RCV	0.751 (0.747–0.755)	0.752 (0.744–0.760)	0.750 (0.750–0.750)
PSA	RF	0.548 (0.545–0.552)	0.103 (0.096–0.110)	0.993 (0.993–0.993)
	DNN	0.784 (0.779–0.790)	0.764 (0.754–0.775)	0.805 (0.805–0.805)
	DPC	0.667 (0.660–0.674)	0.383 (0.369–0.396)	0.951 (0.951–0.951)
	absDPC	0.616 (0.611–0.620)	0.286 (0.277–0.295)	0.945 (0.945–0.945)
	RCV	0.741 (0.737–0.746)	0.863 (0.854–0.872)	0.619 (0.619–0.619)

CI, confidence interval; AFP, alpha-fetoprotein; CA 19-9, carbohydrate antigen 19-9; CA-125, cancer antigen 125; CEA, carcinoembryonic antigen; PSA, prostate-specific antigen; RF, random forest; DNN, deep neural network; DPC, delta percent change; absDPC, absolute delta percent change; RCV, reference change values.

patient class I, DNN performed as well as or worse than RCV, suggesting that model training may have been H or O dominant due to the small proportion of I in the training data (14,223/179,929≈7.9 %). RF's performance varied depending on the tumor marker, with low performance for CA19-9 and PSA compared to other methods.

The difference between ML-based and conventional methods by patient class is noteworthy. Figure 5 shows the difference in balanced accuracy by patient class for each of the five tumor markers. The balanced accuracy tends to decrease for all models in the order of H, O, and I. However, conventional methods show a larger variation, or deviation, by patient class than ML methods. This likely stems from statistical estimation methods' high susceptibility to input data variability, as they detect errors through numbers derived by a fixed methodology. In contrast, ML-based models, which learn potential patterns associated with sample misidentification during the process of training the model regardless of patient class, are more robust to environmental changes than conventional models that rely on numerical variability. In conclusion, the above results suggest that the ML-based DNN model not only performs the best as a sample misidentification error detection model with previous and current test result values as input but is also more robust to clinical environment differences than conventional methods and can be applied flexibly in practical situations. We believe that the results indicate the capability of DNNs to perform effectively when applying delta check models developed based on one manufacturer's device results to another manufacturer's devices. We intend to further investigate and analyze this aspect in future research.

An important consideration in interpreting the results of this study is that it used data generated through *in silico* simulations rather than actual sample misidentification data. This method, which was also applied in Zhou's study [8], can be considered a viable alternative for developing sample misidentification error detection models in current

Table 5: Performance comparison for each delta check model in detecting sample misidentification errors for tumor markers in patient class I (emergency patients or inpatients). Delta check methods based on RF and DNN were developed using a 1 % randomly shuffled total D-set, without regard to patient class. Conventional delta check limits for DPC, absDPC, and RCV were derived using the non-shuffled D-set from the patient class I. Numbers represent the mean (95 % CI).

Tumor marker	Method	Balanced accuracy	Sensitivity	Specificity
AFP	RF	0.684 (0.656–0.711)	0.536 (0.481–0.591)	0.831 (0.831–0.832)
	DNN	0.705 (0.680–0.731)	0.751 (0.700–0.801)	0.660 (0.659–0.660)
	DPC	0.599 (0.577–0.621)	0.229 (0.186–0.273)	0.968 (0.968–0.968)
	absDPC	0.511 (0.500–0.521)	0.050 (0.029–0.072)	0.971 (0.971–0.971)
	RCV	0.619 (0.595–0.644)	0.796 (0.746–0.845)	0.443 (0.443–0.444)
CA19-9	RF	0.596 (0.585–0.607)	0.214 (0.192–0.237)	0.978 (0.978–0.978)
	DNN	0.795 (0.785–0.805)	0.763 (0.743–0.784)	0.826 (0.826–0.827)
	DPC	0.615 (0.604–0.626)	0.273 (0.250–0.295)	0.957 (0.957–0.957)
	absDPC	0.526 (0.520–0.533)	0.104 (0.091–0.118)	0.949 (0.948–0.949)
	RCV	0.690 (0.683–0.698)	0.881 (0.866–0.895)	0.500 (0.500–0.500)
CA125	RF	0.591 (0.575–0.606)	0.238 (0.207–0.268)	0.944 (0.943–0.944)
	DNN	0.742 (0.726–0.758)	0.716 (0.683–0.748)	0.768 (0.768–0.768)
	DPC	0.599 (0.584–0.615)	0.230 (0.199–0.261)	0.969 (0.968–0.969)
	absDPC	0.519 (0.510–0.528)	0.085 (0.068–0.103)	0.953 (0.953–0.953)
	RCV	0.691 (0.674–0.708)	0.724 (0.690–0.758)	0.658 (0.658–0.658)
CEA	RF	0.693 (0.682–0.705)	0.501 (0.478–0.524)	0.886 (0.886–0.886)
	DNN	0.727 (0.717–0.737)	0.757 (0.737–0.777)	0.696 (0.696–0.697)
	DPC	0.623 (0.613–0.634)	0.304 (0.283–0.325)	0.943 (0.943–0.943)
	absDPC	0.542 (0.534–0.550)	0.130 (0.114–0.146)	0.954 (0.954–0.954)
	RCV	0.667 (0.657–0.677)	0.796 (0.776–0.816)	0.538 (0.537–0.538)
PSA	RF	0.529 (0.513–0.544)	0.079 (0.048–0.110)	0.978 (0.978–0.978)
	DNN	0.765 (0.733–0.797)	0.776 (0.712–0.841)	0.753 (0.753–0.754)
	DPC	0.706 (0.674–0.739)	0.469 (0.404–0.534)	0.944 (0.943–0.944)
	absDPC	0.513 (0.495–0.531)	0.089 (0.053–0.124)	0.938 (0.937–0.938)
	RCV	0.650 (0.619–0.681)	0.846 (0.785–0.908)	0.454 (0.453–0.454)

CI, confidence interval; AFP, alpha-fetoprotein; CA 19-9, carbohydrate antigen 19-9; CA-125, cancer antigen 125; CEA, carcinoembryonic antigen; PSA, prostate-specific antigen; RF, random forest; DNN, deep neural network; DPC, delta percent change; absDPC, absolute delta percent change; RCV, reference change values.

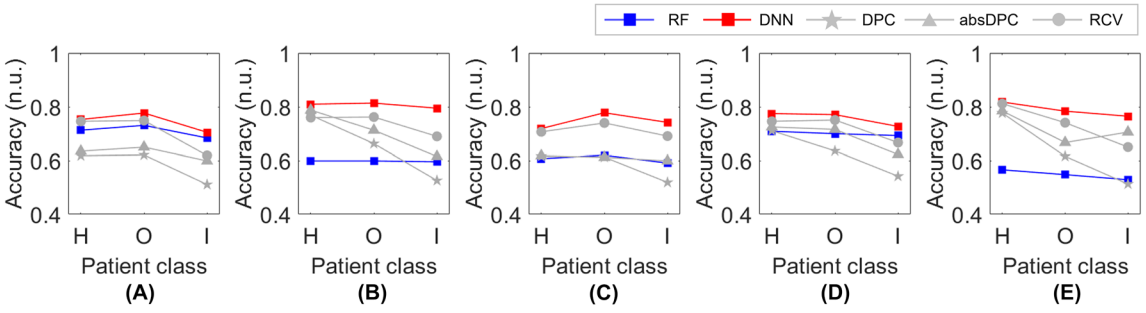

Figure 5: Balanced accuracy by patient class for each of the five tumor markers. (A) AFP, (B) CA19-9, (C) CA125, (D), CEA, (E) PSA. AFP, alpha-fetoprotein; CA 19-9, carbohydrate antigen 19-9; CA-125, cancer antigen 125; CEA, carcinoembryonic antigen; PSA, prostate-specific antigen; RF, random forest; DNN, deep neural network; DPC, delta percent change; absDPC, absolute delta percent change; RCV, reference change values; H, health screening; O, outpatients; I, emergency patients or inpatients.

laboratory testing environments where sample misidentifications and associated test results are not recorded in the LIMS. Collecting real-world sample misidentification data will allow us to understand the different conditions that affect sample misidentification, as well as test results, so that we can improve error detection performance with additional inputs. Moreover, this study provided an important step in confirming the feasibility of ML in detecting tumor

market misidentification errors, however, it focused on internal validation and did not include validation of generalizability. In a follow-up study, we aim to overcome this limitation by conducting external validation using data from various organizations, which is an important step before clinical application.

Conclusions

We developed a delta check method based on RF and DNN models for the five most commonly used tumor markers in clinical laboratories and evaluated it against the conventional delta check method for its performance in detecting sample misidentification errors. The DNN model showed the highest overall values of balanced accuracy considering both sensitivity and specificity for all tumor markers, suggesting that it is the most appropriate delta check method for use in clinical laboratories. In addition, the DNN has shown that it can be applied to multiple patient classes with development with a single total dataset, which is expected to be useful in the future for various checks in clinical laboratories.

Acknowledgments: This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR20C0026), and was supported by a grant (2023IP0003-1) from the Asan Institute for Life Sciences, Asan Medical Center, Seoul, Korea.

Research ethics: The data used in this study was collected after approval by the Ethics Review Boards of each institution, and the requirement for human consent was exempted as a retrospective study (PNUH 2210-023-120, HPIRB 2022-09-017, ISPAIK 2022-09-031).

Informed consent: This study is an analysis using retrospective data. We obtained exemption from informed consent from the IRBs of each institution.

Author contributions: Conceptualization, S.K.; data curation, Y.C., S.Y., and KH.S.; formal analysis, HS S.; funding acquisition, S.K. and H.S.; methodology, HS.S.; supervision, H.S.; writing – original draft, HS.S. and Y.C.; writing – review and editing, S.K and H.S. All authors have read and agreed to the published version of the manuscript.

Competing interests: The authors state no conflict of interest.

Research funding: This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR20C0026), was supported by a grant (2023IP0003-1) from the Asan Institute for Life Sciences, Asan Medical Center, Seoul, Korea, and was supported by a grant from the Sysmex Korea.

Data availability: The IRBs of each hospital did not approve the sharing of raw data (because they are patients' results).

References

1. Desai S, Guddati AK. Carcinoembryonic antigen, carbohydrate antigen 19-9, cancer antigen 125, prostate-specific antigen and other cancer markers: a primer on commonly used cancer markers. World J Oncol 2023;14:4–14.
2. Chang J, Kim S, Yoo SJ, Park EJ, Um TH, Cho CR. Preanalytical errors in the Central Laboratory of a University Hospital based on the analysis of year-round data. Clin Lab 2020;66:1783–91.
3. Lippi G, Chance JJ, Church S, Dazzi P, Fontana R, Giavarina D, et al. Preanalytical quality improvement: from dream to reality. Clin Chem Lab Med 2011;49:1113–26.
4. Lippi G, Cadamuro J, von Meyer A, Simundic AM, European Federation of Clinical C, Laboratory Medicine Working Group for Preanalytical P. Practical recommendations for managing hemolyzed samples in clinical chemistry testing. Clin Chem Lab Med 2018;56:718–27.
5. Clinical and Laboratory Standards Institute. Use of delta checks in the medical laboratory, 2nd ed. Wayne, PA, USA: CLSI guideline EP33; 2023.
6. Ovens K, Naugler C. How useful are delta checks in the 21 century? A stochastic-dynamic model of specimen mix-up and detection. J Pathol Inf 2012;3:5.
7. Yu S, Shin KH, Shin S, Lee H, Yoo SJ, Jun KR, et al. Practical delta check limits for tumour markers in different clinical settings. Clin Chem Lab Med 2023;61:1829–40.
8. Zhou R, Liang YF, Cheng HL, Wang W, Huang DW, Wang Z, et al. A highly accurate delta check method using deep learning for detection of sample mix-up in the clinical laboratory. Clin Chem Lab Med 2022;60: 1984–92.
9. Rosenbaum MW, Baron JM. Using machine learning-based multianalyte delta checks to detect wrong blood in tube errors. Am J Clin Pathol 2018;150:555–66.
10. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020;9:14.
11. Farrell CJ. Identifying mislabelled samples: machine learning models exceed human performance. Ann Clin Biochem 2021;58:650–2.
12. Mitani T, Doi S, Yokota S, Imai T, Ohe K. Highly accurate and explainable detection of specimen mix-up using a machine learning model. Clin Chem Lab Med 2020;58:375–83.
13. Breiman L. Random forests. Mach Learn 2001;45:5–32.
14. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digit Signal Process 2018;73: 1–15.
15. Feng C, Wang H, Lu N, Chen T, He H, Lu Y, et al. Log-transformation and its implications for data analysis. Shanghai Arch Psychiatr 2014;26: 105–9.
16. Jiang Y, Cukic B, Menzies T. Can data transformation help in the detection of fault-prone modules? In: ISSTA '08: international

symposium on software testing and analysis. Seattle Washington: Association for Computing Machinery; 2008.
17. European Federation of Clinical Chemistry. EFLM biological variation database. Available from: https://biologicalvariation.eu/.
18. Wojtalewicz N, Vierbaum L, Kaufmann A, Schellenberg I, Holdenrieder S. Longitudinal evaluation of AFP and CEA external proficiency testing reveals need for method harmonization. Diagnostics 2023;13:2019.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/cclm-2023-1185).