Clinica Chimica Acta 531 (2022) 254–260

Contents lists available at ScienceDirect

Clinica Chimica Acta

journal homepage: www.elsevier.com/locate/cca

A deep learning-based system for assessment of serum quality using sample images

Chao Yang [1], Dongling Li [1], Dehua Sun [1], Shaofen Zhang, Peng Zhang, Yufeng Xiong, Minghai Zhao, Tao Qi, Bo Situ *, Lei Zheng *

Department of Laboratory Medicine, Nanfang Hospital, Southern Medical University, Guangzhou 510515, PR China

ARTICLE INFO

Keywords:
Deep learning
Serum quality
Hemolysis
Icterus
Lipemia

ABSTRACT

Background: Serum quality is an important factor in the pre-analytical phase of laboratory analysis. Visual inspection of serum quality (including recognition of hemolysis, icterus, and lipemia) is widely used in clinical laboratories but is time-consuming, subjective, and prone to errors.

Methods: Deep learning models were trained using a dataset of 16,427 centrifuged blood images with known serum indices values (including hemolytic index, icteric index, and lipemic index) and their performance was evaluated by five-fold cross-validation. Models were developed for recognizing qualified, unqualified and image-interfered samples, predicting serum indices values, and finally composed into a deep learning-based system for the automatic assessment of serum quality.

Results: The area under the receiver operating characteristic curve (AUC) of the developed model for recognizing qualified, unqualified and image-interfered samples was 0.987, 0.983, and 0.999 respectively. As for subclassification of hemolysis, icterus, and lipemia, the AUCs were 0.989, 0.996, and 0.993. For serum indices and total bilirubin predictions, the Pearson's correlation coefficients (PCCs) of the developed model were 0.840, 0.963, 0.854, and 0.953 respectively. Moreover, 30.8% of serum indices tests were deemed unnecessary due to the preliminary application of the deep learning-based system.

Conclusions: The deep learning-based system is suitable for the assessment of serum quality and holds the potential to be used as an accurate, efficient, and rarely interfered solution in clinical laboratories.

1. Introduction

Laboratory activities are virtually partitioned into three phases: pre-analytical phase, analytical phase, and post-analytical phase. The pre-analytical phase encompasses all the procedures before the laboratory analysis, and is responsible for the majority of the laboratory errors [1]. Pre-analytical errors can occur in vivo or in vitro. Major in vivo factors including hemolysis (also occur in vitro), icterus, and lipemia present in diagnostic samples may be regarded as an important pre-analytical factor, which may generate biological, analytical, and physical interference on diagnostic testing [2].

Visual inspection of serum quality is widely used in clinical laboratories, and its accuracy varies among different individuals because of environmental and physiological factors [3]. To reduce subjectivity, colored charts with corresponding degrees of hemolysis are adopted for standardized handling of hemolyzed samples. However a study reported that it still led to an unacceptably high rate (31%) of incorrectly processed hemolyzed samples in the emergency clinical chemistry laboratory [4]. Thus, visual estimation with standardization or not is still subjective and time-consuming.

Some sample pretreatment equipment takes pictures to identify unqualified samples based on the traditional image segmentation algorithm. A sample image is split into two parts and the serum part is used to identify hemolysis, icterus, and lipemia by comparison of the RGB color space. However, this method is susceptible to interference, and studies on the accuracy of this method have not been reported.

Automated detection of hemolysis, icterus, and lipemia using H-index, I-index, and L-index on clinical chemistry analyzers is a more reliable, accurate, and standardized way for assessment of serum quality and is highly recommended [5–7]. But the routine assessment of serum indices (or HIL-indices) will take about 10 min for every sample and may generate suppression of test results which may cause potential harm to

* Corresponding authors.
E-mail addresses: bositu@smu.edu.cn (B. Situ), nfyyzhenglei@smu.edu.cn (L. Zheng).

[1] Chao Yang, Dongling Li, and Dehua Sun contributed equally to this work.

https://doi.org/10.1016/j.cca.2022.04.010
Received 31 December 2021; Received in revised form 9 April 2022; Accepted 9 April 2022
Available online 12 April 2022

emergency patients. Therefore, the application of fast visual recognition from images may reduce the unnecessary serum indices measurement and be a benefit to patients.

Deep learning models especially convolutional neural networks (CNNs) have achieved human-level performance in object-classification tasks and have been widely applied to various clinic medical tasks [8], including recognition of blood cells [9], TBS classification [10], and interpretation of serum protein electrophoresis [11]. But the deep learning-based method for serum quality assessment is still very rare.

Inspired by the above observations, we proposed a new deep learning-based system for automated assessing serum quality to improve the accuracy of visual recognition and reduce the unnecessary serum indices measurement.

2. Method and materials

2.1. Dataset selection

This study took place at the Department of Laboratory Medicine, Nanfang Hospital, Southern Medical University, Guangzhou, China. The study included samples received in the emergency biochemistry laboratory from March 2021 to July 2021. The procedure includes the following steps (Fig. 1). Samples were collected with vacuum blood collection tubes and centrifuged at 1917 g for 10 min to separate serum from blood. Then centrifuged samples were delivered to the Roche pretreatment 612 (P612) and sample images were photographed by the inside camera with a white background panel to eliminate exposure effects. HIL-indices were measured by Roche Cobas 702 (C702) which was connected to P612 by automatic transmission track. The cutoff values of HIL-indices to classify hemolysis, icterus and lipemia were 18 (H-index, SI unit, or 29 for conventional unit), 86 (I-index, SI unit, or 5 for conventional unit), and 40 (L-index, without unit), following the criteria previously developed in our laboratory. The laboratory technician visually assessed collected images and identified the image-interfered samples at last. The image-interfered class was defined as the serum part totally covered by bad rotation, labels, or handwriting. Finally, deep learning models for classification and regression tasks were developed and evaluated.

This study was reviewed by the local ethics committee, and consent was obtained under reference number NFEC-2021-206.

2.2. Network architecture and training

The Inception-Resnet-V2 (Keras 2.2 with TensorFlow 2.4 as the backend.) network was chosen for our tasks [12]. The resolution of sample images was 120x500x3. Training and validation sets were randomly divided at the ratio of 8:2. The models were trained on a workstation with an Intel 9900 k CPU and an NVIDIA GeForce RTX3090 graphics card. The initial learning rate was set as 0.0001 and changed to 1/2 every 10 epochs. The model was trained up to 120 epochs and the weight was saved for the best performance according to the validation loss. Five-fold crosschecks for each deep learning model were implemented and similar performance was achieved (Supplementary Tables 1–3).

2.3. Classification tasks

To evaluate the classification ability of the deep learning model, a classification model to identify qualified, unqualified and image-interfered samples was developed. The unqualified class was defined as HIL-index exceeding the threshold of 18(H-index), 86(I-index), and 40(L-index). As unqualified samples were in the complex situation of combinations of hemolysis, icterus, and lipemia, a model that can output combined results would be more helpful. Therefore, we further implemented a binary classification network with a sigmoid activation function to output the probability of hemolysis, icterus, lipemia, and image-interfered.

The P612 can identify samples into six categories based on the traditional image segmentation algorithm. To evaluate the specific performance of the P612, we planned to train a six-categories deep learning model to compare with it. As a preliminary, 139 hemolytic sample images (H-index > 19), 51 seriously hemolytic sample images (H-index > 44), 137 icteric sample images (I-index > 86), and 117 lipemic sample images (L-index > 40) were collected for the P612 image classification calibration before the dataset collecting. Then, a dataset of 4633 consecutive images was collected as the test dataset to evaluate the P612 and the trained six-categories deep learning model.

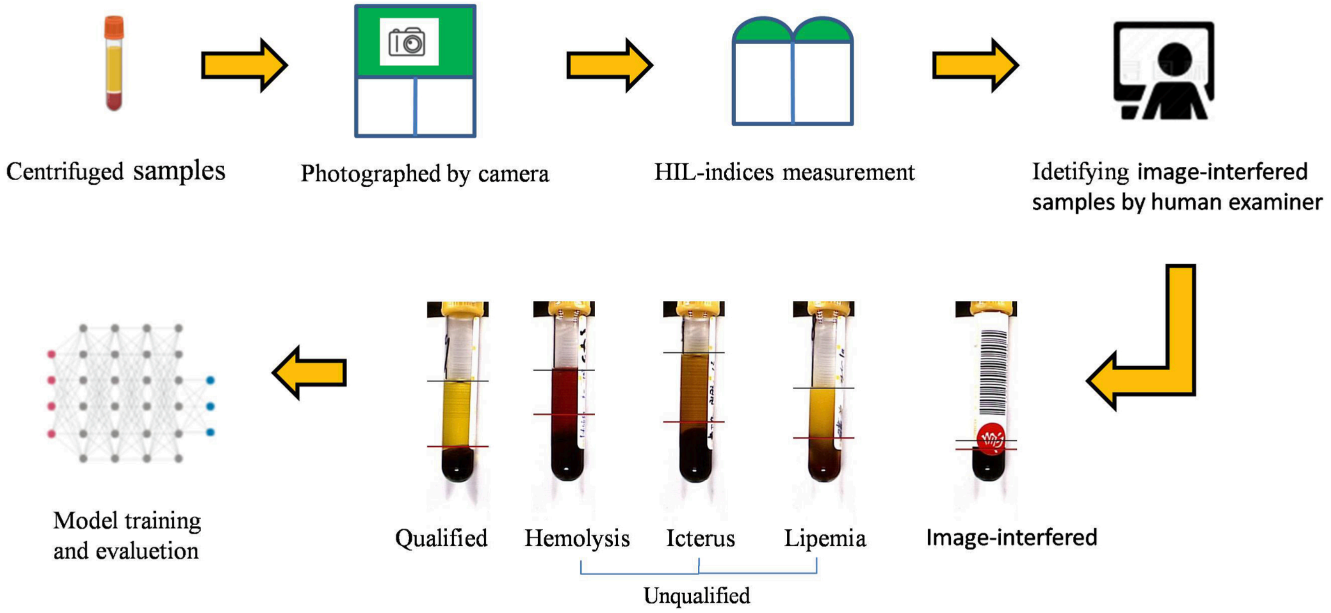

Fig. 1. Data handling workflow. Centrifuged samples were photographed by the camera, HIL-indices results were used to classify hemolysis, icterus, and lipemia. Image-interfered samples were identified by the laboratory technician. Sample images were exported from the LIS and the sample images above exhibited the different status of samples.

2.4. Regression tasks

To evaluate the quantitative prediction ability of the deep learning model, we trained our model to accomplish the regression task (predicting H-index, I-index, and L-index) and chose to use mean squared error (MAE) as a loss function. After achieving the good performance of prediction of HIL-indices, we implemented a further study of prediction of TBIL and TG. First, Samples that tested TBIL and TG by the C702 were obtained from the total dataset. Then study participant metadata including age, sex, TBIL, and TG were collected and analyzed. Finally, a deep learning model for the prediction of TBIL was developed and evaluated.

2.5. Deep learning-based system

We aimed to develop a deep learning-based system that combined image identification with chemical assays for reducing unnecessary HIL-indices tests. The system was developed based on the Roche platform. The deep learning model was developed as a multi-output network. The classification model was used to identify the image-interfered samples. The regression model interacted with LIS was used to determine whether the sample should implement HIL-indices measurement or cancel tests. Then, the test dataset of 4633 consecutive images was used to evaluate the system and another extra dataset was collected after the preliminary application of this system.

3. Results

3.1. Samples

A dataset of 16,427 sample images was collected. Sample images of qualified (68.11%, n = 11190), unqualified (30.98%, n = 5090) and image-interfered (0.89%, n = 147) were shown in Fig. 1. The mean age of the patients was 45.3 ± 18.0, including 9215 men and 7212 women.

Statistics analysis from a continuous dataset (n = 10667, selected from the total dataset.) showed that unqualified samples reached 10.38% of all the biochemical emergency samples. Hemolysis accounted for 61.82% was the main cause of the unqualified samples. Samples with multiple combinations of hemolysis, icterus, and lipemia were shown in Fig. 2a. Hemolytic samples accompanied with lipemia reached 21.88% of the unqualified samples, accounted for 35.39% of hemolytic samples and 66.58% of lipemic samples.

Parts of sample images interfered with an emergency label, handwriting, and patient label (Supplementary Fig. 1). We counted the samples where the serum was partly or totally covered with the interferences in the total dataset. Nearly half (44.63%, n = 7332) of the

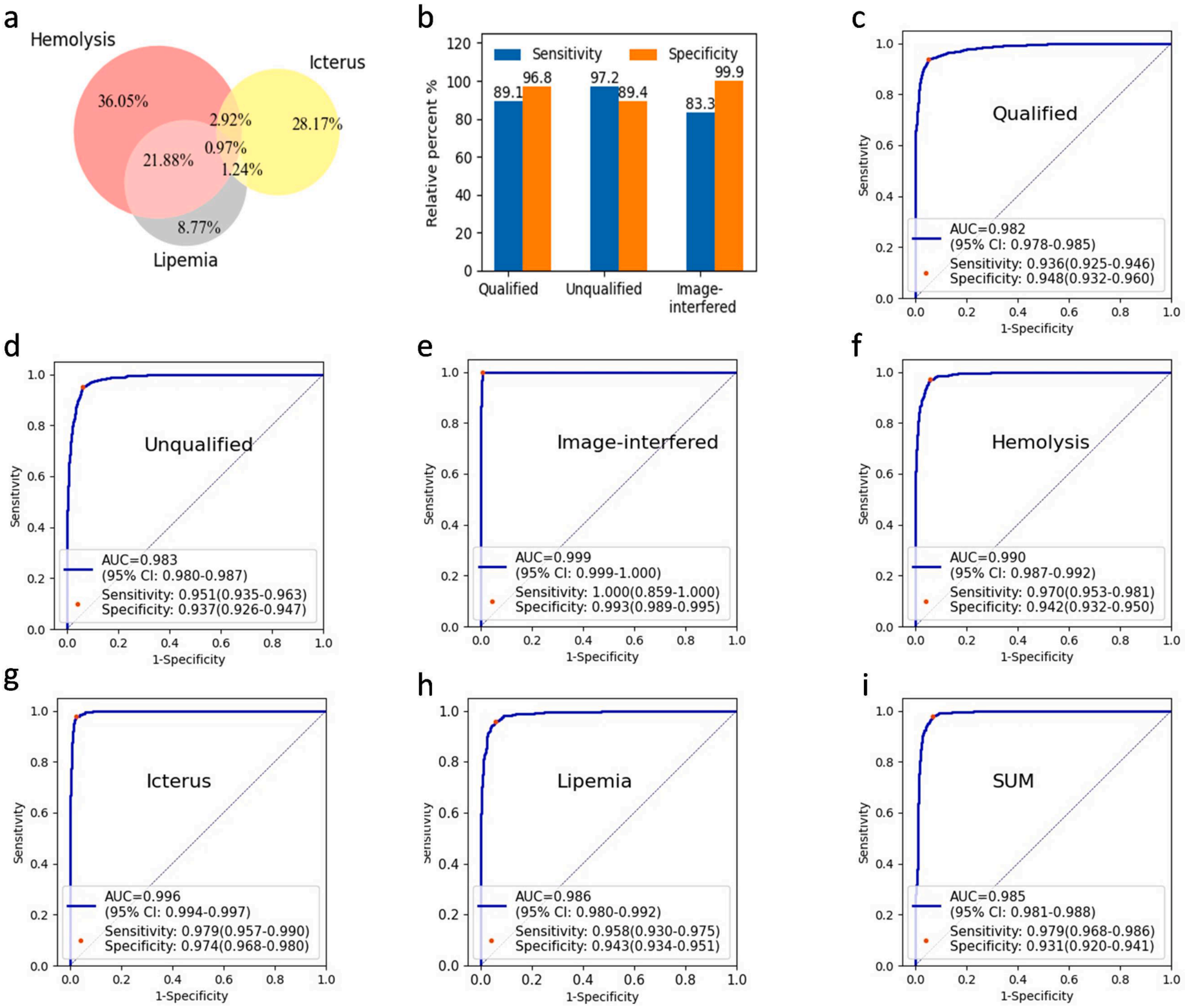

Fig. 2. Performances of deep learning models for classification tasks. (a) Venn chart of component ratio for hemolysis, icterus, and lipemia. (b) Sensitivities and specificities of the three classes were determined by the maximum of output probabilities. (c-e) ROC curves showing performance of classification model in qualified samples, unqualified samples, and image-interfered samples. (f-i) ROC curves showing performance of binary classification models for hemolysis, icterus, lipemia, and P_{SUM}.

sample images were covered with objects in our laboratory (Only images with totally covered serum parts were counted as the image-interfered class). Among them, 1.64% (n = 269) of sample images interfered with an emergency label, 28.03% (n = 4604) interfered with handwriting and 23.31% (n = 3829) interfered with a patient label.

3.2. *Classification of unqualified and image-interfered samples*

The sensitivity and specificity for recognizing unqualified samples were 97.2% and 89.4% (Fig. 2b). The AUC for qualified class was 0.981 (95% CI: 0.978–0.985), unqualified class was 0.983 (95% CI: 0.976–0.986) and image-interfered class was 0.999 (95% CI: 0.998–1.000). To achieve a better sensitivity of the model, we recalculated the specificity of this model. When sensitivity > 95%, sensitivity > 99% and sensitivity = 100%, 6.26%, 18.47% and 48.14% qualified samples will be misclassified to unqualified samples.

3.3. *Binary classification of hemolysis, icterus, and lipemia*

The model achieved good performance for classes of hemolysis, icterus and lipemia with AUCs > 98% (Fig. 2f-h). Also to distinguish unqualified samples with only one network, we determined the sum of each class probability: $P_{\mathrm{SUM}} = P_{\mathrm{hemolysis}} + P_{\mathrm{iterus}} + P_{\mathrm{lipemia}}$. The AUC of the P_{SUM} was 0.986 (Fig. 2i), as well as the aforesaid model in identifying unqualified samples (AUC = 0.983). When class thresholds were set to receive the max AUCs with $t_{\mathrm{hemolysis}} > 0.016$, $t_{\mathrm{iterus}} > 0.001$and $t_{\mathrm{lipemia}} > 0.004$, the sensitivity and specificity of identifying unqualified samples was 0.982 and 0.865 respectively, achieving a comparable performance of P_{SUM} (sensitivity = 0.979, specificity = 0.931). To know which part of the sample image determined the classification results. Saliency maps were generated and shown in Supplementary Fig. 2.

3.4. *Comparison of the deep learning model and the traditional image segmentation algorithm*

For the P612 classification results in the test dataset, 75.61% (n = 3024) of qualified samples were misclassified to other classes, which accounted for 65% of the test dataset. Among them, 40.46% (n = 1640) were misclassified to lipemic samples. 27.73% (n = 1124) were misclassified to icteric samples. 5.57% (n = 226) were misclassified to unknown samples. While for the deep learning model, the overall accuracy was greatly improved from 30.19% to 95.85%. Comparing the confusion matrix plot of the P612 (Fig. 3a), very good performance of the deep learning model (Fig. 3b) is observed for all classes. The full statistics of prediction quality are provided in Table 1.

3.5. *Prediction of HIL-indices*

A linear correlation between the predicted HIL-indices and measured HIL-indices showed strong associations with the coefficient of determination (R^2) of 0.689, 0.928, 0.720, Pearson’s correlation coefficient (PCC) of 0.840, 0.963, 0.854 and mean Square Error (MAE) of 4.25, 8.44, 7.63 (Fig. 4a-c). The level of agreement between the predicted HIL-indices and measured HIL-indices was assessed using a Bland–Altman plot. The deep learning model achieved an intraclass correlation coefficient (ICC) of 0.83 for H-index, 0.96 for I-index, and 0.82 for L-index (Fig. 4d-f). The deep learning model estimated the HIL-indices level to a greater extent at high levels than at low levels.

3.6. *Prediction of TBIL*

10,484 samples that tested TBIL by the C702 were obtained from the total dataset and 1433 samples that tested TG were obtained. The mean age of the patients was 48.1 ± 20.0, including 6349 men and 4135 women. The I-index value showed a high correlation with TBIL (PCC = 0.998, P < 0.01, Fig. 4g), while the PCC of L-index versus TG was 0.764 (P < 0.01, Supplementary Fig. 3c). Other correlation analyses for DBIL, IBIL, CHOL, HDL, and LDL were exhibited in Supplementary Fig. 3a-f. The PCC of predicted TBIL versus measured TBIL was 0.953, the R^2 was 0.907, and the MAE was 8.808 (Fig. 4h) on the validation dataset (n = 2096). As shown in the Bland–Altman plot (Fig. 4i), the deep learning model estimated the TBIL level to a greater extent at high levels than at

Table 1
Class-wise sensitivity and specificity of the Roche P612 and the deep learning model.

Class	Roche P612		Deep learning model		No. of images
	Sensitivity	Specificity	Sensitivity	Specificity	
Qualified	0.253	0.993	0.974	0.882	4053
Hemolysis	0.312	0.993	0.618	0.983	262
Lipids	0.666	0.622	0.755	0.994	180
Bilirubin	0.709	0.735	0.777	0.995	148
Unknown	0.892	0.939	0.678	1.0	56
Severe hemolysis	0.303	0.991	0.449	0.998	89

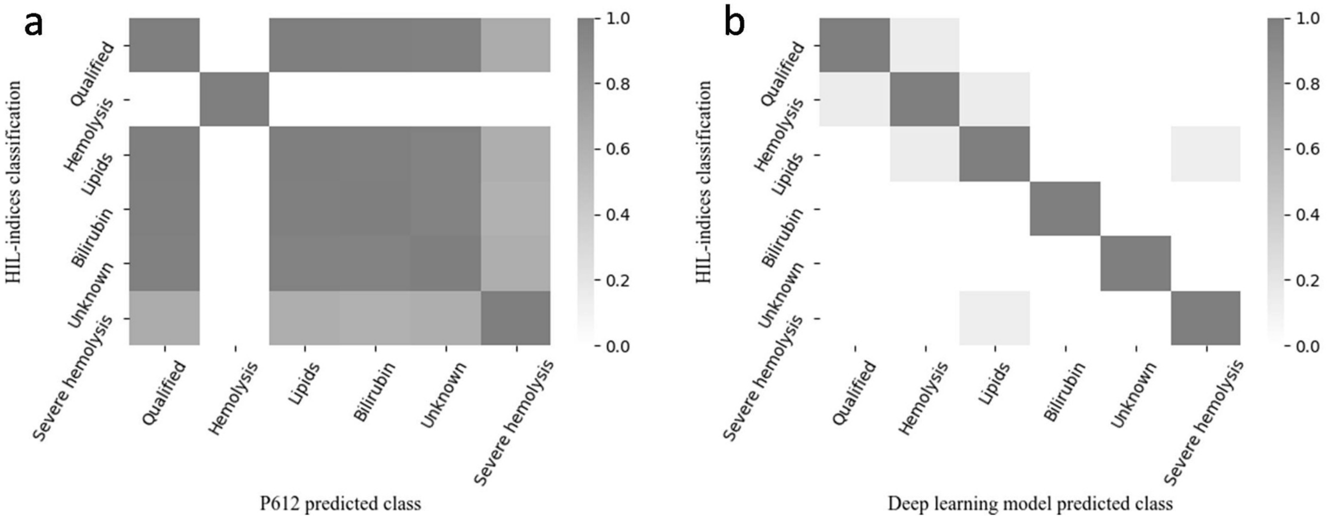

Fig. 3. Comparison of the P612 and the deep learning model. (a) Confusion matrix plot for the P612. The darkness presents the relative frequency of each class. (b) Same as (a), but for the deep learning model.

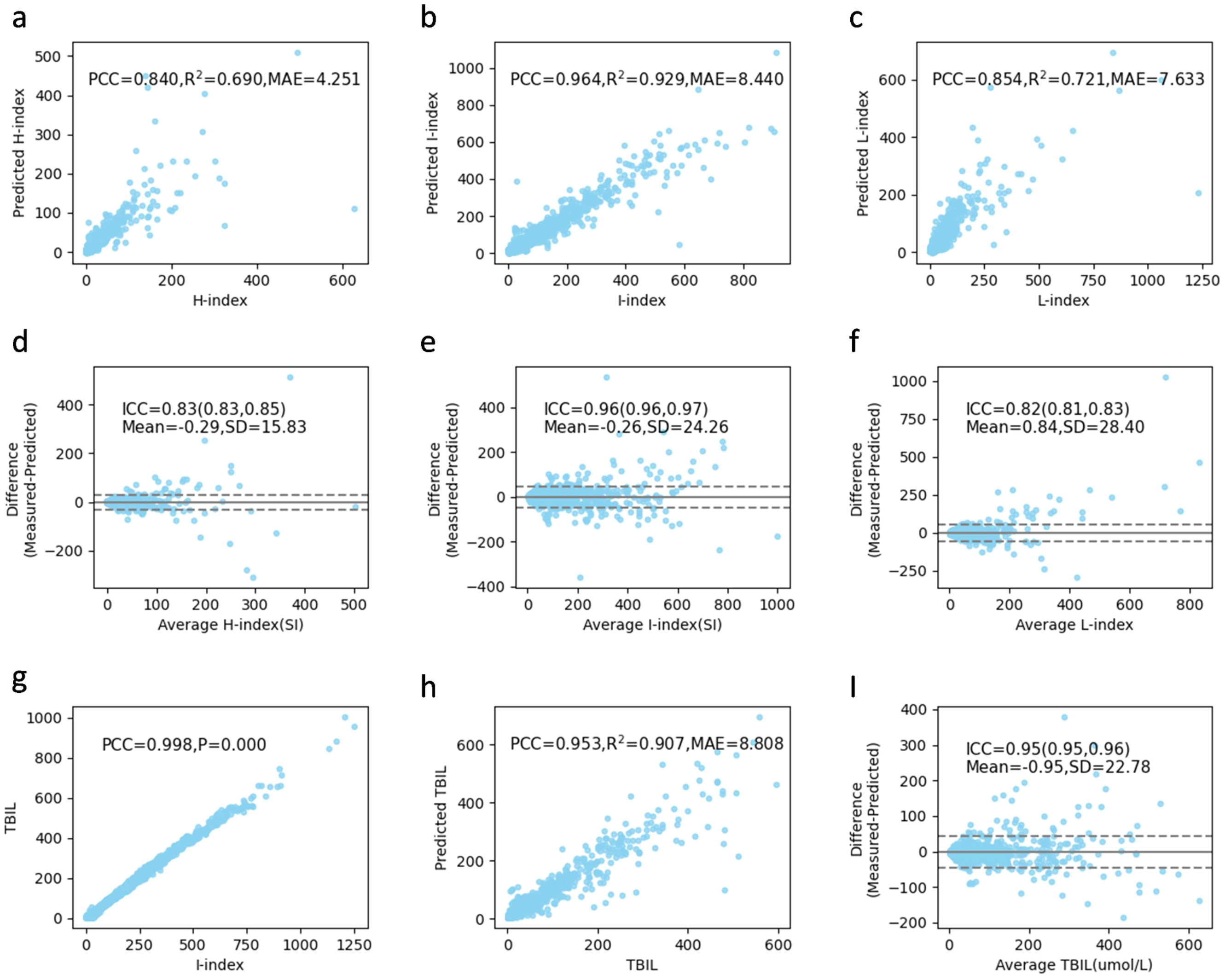

Fig. 4. Performance of deep learning models for HIL-indices and TBIL prediction. (a-c) Scatter plots of HIL-indices prediction. (d-f) Bland–Altman plots of HIL-indices prediction. (g) Correlation analysis of I-index versus TBIL. (h) Scatter plot of TBIL prediction. (i) Bland–Altman plot of TBIL prediction.

low levels. What's more, the deep learning model showed comparable excellent performance in predicting I-index and TBIL.

3.7. *Deep learning-based system for automated assessing serum quality*

The improved workflow of automated assessing serum quality in our laboratory was shown in Fig. 5. After the centrifugal sample was photographed by the camera, the deep learning model received the sample image for identifying and sent the results to LIS. Once the LIS received the results, it would send messages to the instrument to append serum indices or cancel tests. Determined by the classification results, the image-interfered samples were sent to the chemical analyzer and performed HIL-indices tests additionally, while the other samples need to be further processed with the quantitative predicted HIL-indices values. The lower and upper limit for each analyte was set as C-2.58SD and C + 2.58SD respectively (C is the cutoff values for different analytes adopted from the manufacturer's declarations. SD is the standard deviation of the difference between predicted and true values in the test set.). When the predicted HIL-indices values were greater than the upper limit, the interfered test will be canceled. When the predicted HIL-indices values were lower than the lower limit, the sample can be reported normally. When the predicted HIL-indices values were between the lower limit and the upper limit, The HIL-indices tests were appended.

Serum indices tests were reduced for 26.76% (n = 1225) of samples on the test dataset (n = 4577, excluded 56 image-interfered samples). Among the 26.76% of samples, 1 sample was misclassified for predicted serum indices lower than the lower limit but measured serum indices greater than the cutoff values, 3 samples were misclassified for predicted serum indices greater than the upper limit but measured serum indices lower than the cutoff values. As shown in Supplementary Tabel 4, the lower rate of reducing HIL-indices assays was mainly due to the low cutoff value of the H-index for HBDH, LDH, and AST. Further steps to defining cutoffs for flagging, alarming, or suppressing test results with predicted HIL-indices values in our laboratory should be taken and more samples may avoid HIL-indices tests expectantly.

3807 sample images were collected from September 2nd to September 8th after the primary application of the system. 30.8% of samples avoid HIL-indices assays.

4. Discussion

As shown in Fig. 2a, up to 66.58% of lipemic samples, meanwhile, accompanied with hemolysis. Consulting to the reagent instructions, L-index was determinate by a simple measure of two wavelengths of 660 nm and 700 nm. However, the lipemia has a broad-spectrum absorption peak which covered that of hemolysis (at 570 nm and 600 nm) and icterus (at 480 nm and 505 nm). Thus, further studies aimed at dissecting overlapping interference for serum indices are still needed.

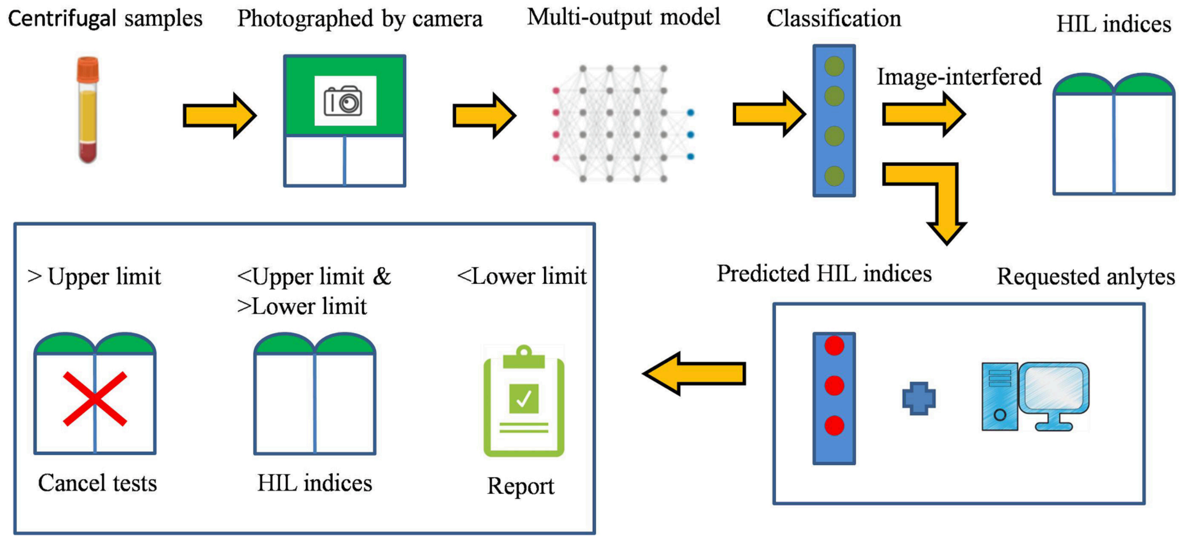

Fig. 5. The improved workflow of the deep learning-based system for automated assessing serum quality. The system was developed as a multi-output network. The classification results were used to identify the image-interfered samples. The regression results interacted with LIS were used to determine whether the sample should implement HIL-indices measurement.

The deep learning models presented in this study show outstanding performance on assessing serum quality with sample images under the situation that nearly half of the images interfered with objects. The models attain the AUCs above 98% in the classification of hemolytic, icteric, lipemic, and image-interfered samples. It offers the promise that more accurate models can be achieved once the image interferences are avoided and sharper cameras are being equipped. Considering the complex situation in different laboratories, it is recommended that every laboratory should collect their own sample images and train deep learning models to improve the accuracy of image recognition and it is the right time to discard the unstable and ineffective human visual check.

Holding unqualified samples for analysis can quickly delay a result to unacceptable lengths of time that can potentially cause more harm to the patient. With the HIL-indices prediction using sample images, we can discover the problematic sample and take steps instantly, especially for emergency patients. Our model shows over 0.84 PCCs of predict HIL-indices versus measured HIL-indices and the PCC of I-index reached up to 0.963. It provides the basis for further research in optimizing HIL-indices tests workflow and a new method for HIL-indices quality control [2].

Hyperbilirubinemia is a manifestation of underlying liver and biliary tract disease. An acute jaundice onset is extremely important and raises the possibility of an acute process such as hepatitis, or complete biliary tract obstruction [13]. Delayed diagnosis of biliary atresia is an important cause of pediatric end-stage liver failure and liver transplantation [14]. Our deep learning model shows a high PCC of 0.95 in the prediction of TBIL using sample images. It can be expected to be developed as a new fast and easy method for TBIL measurement and can be applied in clinical practice.

Compare with Roche P612, our model shows a greater anti-interference capability and can output not only classification results but also regression results. With the rapid development of artificial intelligence, deep learning or other new algorithms will finally replace the traditional image segmentation algorithm and provide more potential applications. Further research in transferring the ability of deep learning models to detect sample clock and serum volume is achievable.

In conclusion, our method holds the potential to act as a rapid and efficient solution for automated assessment of serum quality and might further be applied in more laboratories and achieve better performance.

Author contributions

Chao Yang: designed research, performed research, designed algorithm, and wrote the paper.
Dongling Li: performed research and wrote the paper.
Dehua Sun: performed research and wrote the paper.
Shaofen Zhang: established dataset and analyzed data.
Peng Zhang: established dataset.
Yufeng Xiong: analyzed data and wrote the paper.
Minghai Zhao: established dataset.
Tao Qi: established dataset.
Bo Situ: designed research, performed research, and wrote the paper.
Lei Zheng: designed research, performed research, and wrote the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors acknowledge the support received from medical workers in Department of Laboratory Medicine, Nanfang Hospital, Southern Medical University.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cca.2022.04.010.

References

[1] G. Lima-Oliveira, W. Volanski, G. Lippi, G. Picheth, G.C. Guidi, Pre-analytical phase management: a review of the procedures from patient preparation to laboratory analysis, Scand. J. Clin. Lab. Invest. 77 (3) (2017) 153–163.
[2] G. Lippi, J. Cadamuro, A. von Meyer, A.-M. Simundic, Local quality assurance of serum or plasma (HIL) indices, Clin. Biochem. 54 (2018) 112–118.
[3] G. Lippi, J. Cadamuro, Visual assessment of sample quality:quo usque tandem ? Clin. Chem. Lab. Med. (CCLM) 56 (2018) 513–515.
[4] A.H. Luksic, N. Nikolac Gabaj, M. Miler, L. Dukic, A. Bakliza, A. Simundic, Visual assessment of hemolysis affects patient safety, Clin. Chem. Lab. Med. (CCLM) 56 (2018) 574–581.

[5] G. Lippi, J. Cadamuro, A. von Meyer, A. Simundic, F.O.C.C. European, Practical recommendations for managing hemolyzed samples in clinical chemistry testing, Clin. Chem. Lab. Med. 56 (2018) 718.

[6] A.-M. Simundic, G. Baird, J. Cadamuro, S.J. Costelloe, G. Lippi, Managing hemolyzed samples in clinical laboratories, Crit. Rev. CL Lab. Sci. 57 (1) (2020) 1–21.

[7] G. Lippi, P. Avanzini, D. Campioli, G. Da Rin, M. Dipalo, R. Aloe, D. Giavarina, G. L. Salvagno, Systematical assessment of serum indices does not impair efficiency of clinical chemistry testing: A multicenter study, Clin. Biochem. 46 (13-14) (2013) 1281–1284.

[8] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nat. Med. 25 (1) (2019) 24–29.

[9] C. Matek, S. Schwarz, K. Spiekermann, C. Marr, Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks, Nat. Mach. Intell. 1 (11) (2019) 538–544.

[10] X. Zhu, X. Li, K. Ong, W. Zhang, W. Li, L. Li, D. Young, Y. Su, B. Shang, L. Peng, W. Xiong, Y. Liu, W. Liao, J. Xu, F. Wang, Q. Liao, S. Li, M. Liao, Y.u. Li, L. Rao, J. Lin, J. Shi, Z. You, W. Zhong, X. Liang, H. Han, Y. Zhang, N.a. Tang, A. Hu, H. Gao, Z. Cheng, L.i. Liang, W. Yu, Y. Ding, Hybrid AI-assistive diagnostic model permits rapid TBS classification of cervical liquid-based thin-layer cell smears, Nat. Commun. 12 (1) (2021), https://doi.org/10.1038/s41467-021-23913-3.

[11] F. Chabrun, X. Dieu, M. Ferre, et al., Achieving Expert-Level Interpretation of Serum Protein Electrophoresis through Deep Learning Driven by Human Reasoning, Clin. Chem. (2021).

[12] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv e-prints, 2016, pp. 1602–7261.

[13] J.I. Sullivan, D.C. Rockey, Diagnosis and evaluation of hyperbilirubinemia, Curr. Opin. Gastroen. 33 (2017) 164–170.

[14] L. Lam, S. Musaad, C. Kyle, S. Mouat, Utilization of Reflex Testing for Direct Bilirubin in the Early Recognition of Biliary Atresia, Clin. Chem. 63 (2017) 973–979.