Reproducing Human Subjective Evaluation in the Microscopic Agglutination Test with Deep Learning

Risa Nakano[1], Yuji Oyamada[1*], Ryo Ozuru[2*], Satoshi Miyahara[3], Michinobu Yoshimura[2], Kenji Hiromatsu[2]

[1]Graduate School of Engineering, Tottori University, Tottori, Japan

[2]Department of Microbiology and Immunology, Faculty of Medicine, Fukuoka University, Fukuoka, Japan

[3]Department of Microbiology, School of Medicine, University of Occupational and Environmental Health, Kitakyushu, Japan

[*] Addresses correspondence to Yuji Oyamada, oyamada@tottori-u.ac.jp, Ryo Ozuru, ozuru@fukuoka-u.ac.jp.

ABSTRACT

The Microscopic Agglutination Test (MAT) is widely recognized as the gold standard for diagnosing zoonosis leptospirosis. However, a significant limitation of MAT is the inconsistency in test results, as it relies on the examiners' subjective perceptions to estimate agglutination rates. To address this issue, we propose a deep neural network to replicate the subjective evaluation process of agglutination rate estimation in MAT. By employing a pre-trained DenseNet121, we can efficiently optimize the network parameters during the training phase. We validated the trained network using our

original dataset. Experimental results demonstrate that the proposed network provides accurate agglutination rate estimates. Furthermore, we utilize a standard visualization technique to gain insights into the decision-making process of the deep learning methods. The findings reveal that the proposed network extracts image features indicative of leptospire abundance. Overall, these results suggest that deep learning is effective for estimating agglutination rates and that enhancing interpretability aids medical experts in understanding the functionality of deep learning.

INTRODUCTION

Leptospirosis is a zoonosis caused by the pathogenic bacterium *Leptospira* [1]. It is considered one of the neglected tropical diseases due to its prevalence in tropical and subtropical regions with high rainfall or typhoons. Yet, it receives less attention compared to other infectious diseases like malaria and AIDS [2]. Preventing the spread of leptospirosis requires serovar-specific vaccination, so it is essential to identify the serovars prevalent in the region from the over 250 recognized serovars.

The Microscopic Agglutination Test (MAT) is the standard test for serovar identification and definitive diagnosis of leptospirosis by indirectly estimating the amount of antibodies in patient sera based on the strength of antigen-antibody reactions [3, 4]. The antigen-antibody reaction occurs only when the test serum is mixed with its corresponding serovar of leptospires as antigens. More antibodies bind more antigens, and agglutination formed by the binding of antigens precipitates over time. Less agglutination is formed with a more diluted test serum because dilution reduces the amount of antibodies, resulting in weak antigen-antibody reactions. In MAT, we estimate the reduction of free leptospires by adding test serum as an indirect estimation of the strength of antigen-antibody reactions.

The MAT procedure is based on the above characteristics. Test serum is mixed with various serovars of leptospires in a serial dilution. For each dilution factor, we measure the strength of the antigen-antibody reaction by comparing the mixed solution against the negative control (no-sera control). The measured strength ranges from 0 to 100% and is

determined subjectively. MAT has several drawbacks. Subjective evaluation heavily relies on laboratory technicians' proficiency and the evaluation criteria may differ depending on the technicians. In addition, the test is time-consuming because the procedures must be repeated for each candidate serovar.

As a first step in the computerization of MAT, we have researched machine learning adaptation of MAT based on traditional machine learning techniques [5]. The proposed method uses binary classification to categorize a dark-field microscopy image as either agglutinated or not agglutinated based on the appearance of free leptospires. While this is a worthwhile study as the first application of machine learning to MAT, there are two areas for improvement. The extracted features cannot distinguish free leptospires from other obstacles of similar size. In addition, the method ignores step dilution, and thus, highly diluted test serum may be classified as agglutinated even when the test serum is mixed with the matched serovar leptospires.

Deep learning [6] has been actively introduced in medical image processing [7]. In contrast to traditional machine learning techniques, deep learning techniques are constructed on a single framework, allowing researchers can concentrate on designing network architecture and optimizing its parameters. The network architecture represents how networks obtain output from input data, effectively mimicking the cognitive processes of medical experts. The network parameters represent the internal variables that the network uses for its computations. Appropriate parameters are obtained through the training process.

This paper aims to build a deep-learning method for calculating the agglutination rate of MAT. For this purpose, we propose a deep neural network that mimics MAT's subjective evaluation. Furthermore, we utilize a visualization technique that gives us clues as to why and how deep learning methods make their decisions. This technique is frequently used when deep learning is applied for medical purposes [8]. Please note that all the experiments in this paper were conducted in a laboratory. Experiments and verification with diverse institutes are outside the scope of this paper.

MATERIALS AND METHODS

Acquisition of MAT image data

MAT was performed using standard methods with rabbit sera obtained in the previous study [9] We studied the following *Leptospira* serovars: Poi, Losbanos, Manilae, and Icterohaemorrhagiae. For each test serum, we performed MAT with one matched and one unmatched serovar. Table 1 shows the eight serovar-serum combinations. The test sera were diluted in 2-fold increments from 1:10 to 1:10,240 and incubated at 30°C for 3 hours after *Leptospira* was added so that the final concentrations were $\sim 10^7$/mL. Each sample was placed on a glass slide followed by a cover glass and observed under a dark-field microscope (BX-50, Olympus, Tokyo, Japan) using a CCD camera (Basler ace acA800-510um, Basler AG, Ahrensburg, Germany). We took 2,880 ($8 \times 12 \times 30$) photographs, 30 from different fields of view per sample. All photographs were saved in TIFF format with a resolution of 512×512. Hereafter, the 11 images with different dilution magnifications are called *inspection images*. On the other hand, photographs capturing the negative control, which is no-sera control, are called *reference images*.

TABLE 1 Experimental serovar-serum combinations.

Matched/Unmatched	Case name	Serovar	Test serum
Matched cases	Poi-Poi	Poi	Poi
	Los-Los	Losbanos	Losbanos
	Man-Man	Manilae	Manilae
	Ict-Ict	Icterohaemorrhagiae	Icterohaemorrhagiae
Unmatched cases	Poi-Ict	Poi	Icterohaemorrhagiae
	Los-Poi	Losbanos	Poi
	Man-Los	Manilae	Losbanos
	Ict-Man	Icterohaemorrhagiae	Manilae

Data

Here, the following processing is applied to the photos to create a dataset for the experiments.

First, we created 79,200 image pairs from 2,880 images of the eight serovar-serum combinations. For each serovar-serum combination, we made 9,900 image pairs by coupling 30 reference images with 330 inspection images.

Next, we annotated the agglutination rates for each pair. The annotated agglutination rates were assigned to pairs of the same dilution factor. The same agglutination ratio was assigned to multiple pairs for two reasons: ideally, the amount of agglutination observed in photographs taken on the same glass slide would be perfectly consistent, and to reduce the cost of annotation. The annotated values were set to 11 levels, ranging from 0 to 100% in 10% intervals. The quantization levels were determined based on actual test values, but there is no medical or laboratory basis. Table 2 shows the annotated values for each dilution factor; of the eleven annotation values, seven were used (0, 20, 30, 50, 70, 90, and 100). These annotated values are used as the ground truth for the experiments. When the serovar and test sera were correctly matched, the agglutination rate decreased as the dilution factor increased. In the case of mismatched servos, the agglutination ratio was almost zero regardless of the dilution factor. The Ict-Man pair shows a high agglutination rate at low dilution factors despite the serovar-serum mismatch. This suggests the possibility of cross-reaction. Notably, this tendency may be attributed to individual differences, and there is no assurance that the Ict-Man pair will consistently exhibit such reactions.

TABLE 2 Serum type-annotated values of agglutination rate for each serum tested. These values are used as the ground truth in the experiments.

Dilution factor	10	20	40	80	160	320	640	1280	2560	5120	10240
Poi-Poi	90	90	100	100	100	100	100	100	70	50	50
Los-Los	70	90	90	100	100	90	70	50	30	0	0
Man-Man	70	90	90	100	100	100	100	100	100	100	100
Ict-Ict	70	70	90	100	100	100	100	90	70	30	0
Poi-Ict	20	0	0	0	0	0	0	0	0	0	0
Los-Poi	0	0	0	0	0	0	0	0	0	0	0
Man-Los	0	0	0	0	0	0	0	0	0	0	0
Ict-Man	90	90	70	30	20	0	0	0	0	0	0

Finally, the 79,200 MAT image pairs were divided into subsets for training and evaluating the proposed neural network. The training dataset optimized the network parameters and UMAP, while the test dataset was used to evaluated the trained network.

Proposed deep neural network

Here we describe our proposed network and experiments to verify its accuracy.

The input data consisted of a MAT image pair: an inspection image and its corresponding negative control as a reference image. The proposed deep neural network estimated the agglutination rate by comparing the inspection image to the reference image, expressed as a percentage. The network comprises three sub-networks: a feature extraction network, a feature comparison network, and a regression network, as shown in Fig. **1**.

The feature extraction network independently extracts discriminative image features from the input images. It is a twin neural network with the same architecture and weights, where each branch processes one of the paired images. We used the feature extraction part of DenseNet121 [10] and added a global pooling layer at the end. This architecture

produces a 1024-dimensional vector as the image feature output. These image features are intended to represent the number of bacteria visible in the input images, though each vector element does not directly correspond to a physical quantity.

The feature comparison network calculates a pair feature by subtracting the extracted image features. This approach assumes that image features are distributed according to the number of bacteria in the feature space.

The regression network estimates the agglutination rate of the image pair from the calculated pair feature, outputting it as a real percentage. It is a feedforward neural network with three fully connected layers separated by ReLU activation functions. The network processes the 1024-dimensional pair feature through two hidden layers with 256 and 64 units, respectively, and outputs an agglutination rate ranging from 0 to 100%.

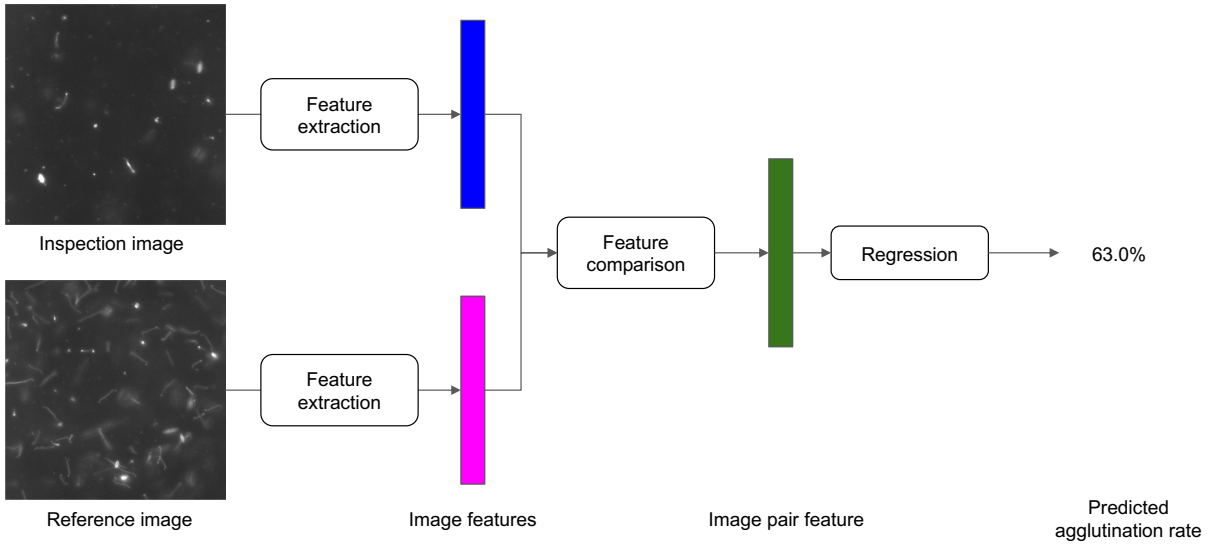

FIG 1 The architecture of the proposed deep neural network.

Network parameters optimization

This section describes how we trained the network parameters. We first explain the training flow and then describe the settings of the training algorithm.

For the network parameter optimization, we took several training steps as shown in Table 3 shows how each sub-networks optimized. As mentioned above, the feature extraction network was pre-trained for natural image classification while the regression network was designed for our problem and initialized by He's method [11]. This means that the feature extraction network was optimized for natural image classification while the regression network was not optimized for any purposes in the initial state. In step 1, we optimized only the regression network so that the entire network can estimate the agglutination rate from natural image features. In step 2, we optimized only the feature extraction network so that the network can extract MAT image features. Finally, we optimized the entire network so that the network can extract MAT image features suitable for agglutination rate estimation and can estimate the agglutination rate from the extracted features.

TABLE 3 The optimization steps. Which type of images each sub network is optimized in each optimization step.

	Feature extraction network	Regression network
Initial state	Natural images	None
Step 1	Natural images	Natural images
Step 2	MAT images	Natural images
Step 3	MAT images	MAT images

Adam is used as the training algorithm for all three training steps. The learning rates were set to 1.0×10^{-3} when learning parameters of the regression network from the initialization state, while the learning rate for subsequent learning was set to 1.0×10^{-5} to keep the parameters fine-tuned. The number of epochs was set to 50 so the learning error would converge during the learning phase. The learning error is the absolute squared error between the ground truth, which is shown in Table 2, and its estimate. We also applied data augmentation to rotate each image at each epoch randomly.

Agglutination rate estimation experiment

We used the test dataset to evaluate the performance of the trained network. For each image pair of the dataset, we compared the annotated agglutination rates with the estimated values.

Visualization

In this section, we describe a visualization technique to explain the rationale behind the decisions of the proposed network. As previously mentioned, it is often challenging to interpret the feature vectors and comparison vectors derived by our proposed networks, as these do not directly correspond to physical quantities. To address this challenge, we introduced UMAP visualization [12], a common technique in computer-aided diagnosis (CAD) when deep learning networks are employed. UMAP converts high-dimensional vectors into corresponding two-dimensional spaces while preserving the global structure of the feature distribution. By coloring the distribution in the converted two-dimensional spaces based on various attributes, we aim to identify important attributes for the proposed network. This visualization aids in providing a clearer understanding of how our proposed network perceives differences in individual images and image pairs.

Visualization of image features

For the visualization of feature vectors, we used only reference images and colored each converted vector based on its serovar-serum pair. Feature vectors are intended to capture the amount of agglutination, and therefore, negative control must show no difference in the feature space.

Visualization of image pair features

For feature comparison vector visualization, we used all the matched image pairs and colored each feature vector based on its agglutination rate. All the unmatched image pairs are omitted from this visualization because, in theory, reference and inspection images

should show no difference. The agglutination rate was used as the coloring metric because feature comparison vectors are intended to capture the difference in agglutination rates while remaining unaffected by other factors.

RESULTS

In this section, we present the results and discussions obtained from the experiments described in the previous chapter, as well as feedback from microbiology and infectious diseases experts. The annotated values shown in Table 2 are used as the ground truth in these experiments. The estimation error is defined as the absolute difference between a predicted value and its corresponding ground truth.

Experiment on Agglutination rate estimation

Fig. 2 compares the ground truth agglutination rate annotated by a laboratory technician with the estimated rates. Fig. 2.A shows the statistics of the estimated values relative to the true values as a box-and-whisker diagram. It indicates a positive correlation between the true values and the estimated values. As the true value increases, the estimation error (the absolute difference between an estimated value and its corresponding ground truth) also tends to increase. The data in the boxed portion of the box-and-whisker diagram (from the second to the third quartile) are solid. Experts provided feedback indicating that they were pleased with the quantification and visualization of the agglutination rate, and noted that the error between the true value and the estimated value was about 10%, which is an acceptable error.

Fig. 2.B shows the frequency and bias of the estimated values relative to the true value. It demonstrates that data with a true value of 0, which accounts for about half of the data, can be accurately estimated. However, data with a true value of 20 or more tend to have lower estimated values than the true value, especially for data with the highest frequency (50 or more), where the predicted agglutination rate is about 20% lower.

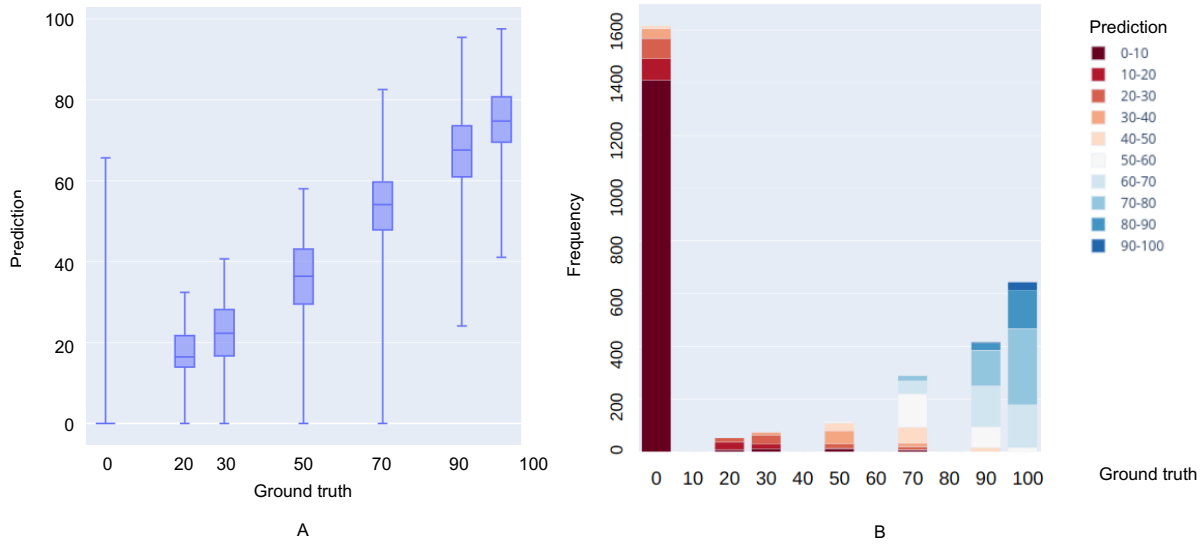

FIG 2 Agglutination rate estimation experiment. A: Boxplot showing the correlation between the annotated agglutination rate and the estimated agglutination rate. B: Stacked histogram showing the frequency of the predicted agglutination rate against the annotated agglutination rate.

The experts' feedback supports our claim that the proposed network can estimate the agglutination rate with a certain degree of accuracy. However, since some data have significant errors, it is necessary to analyze which types of image pairs cause significant errors.

The primary causes of the larger error are the imbalance of the dataset and the imperfect annotation. As shown in Table 2, our dataset is highly imbalanced, with roughly half of the dataset having a 0% agglutination rate. We trained the network by treating all image pairs equally, which resulted in less accurate estimates for higher agglutination rate data.

Next, Fig. 3 shows the change in the agglutination rate with the dilution factor. The positive cases, shown with red borders, maintain a high agglutination rate even at higher dilution factors. On the other hand, the negative cases, shown with blue borders, maintain a low agglutination rate regardless of the dilution factor. In the case of Ict-Man, the agglutination rate is higher only at lower dilution factors, indicating the occurrence of

cross-reaction. While these results are based on the limited amount of data, it is noteworthy that the proposed network effectively visualizes the difference between the cross-reaction and positive infection.

The results imply that the proposed network can distinguish negative, positive, and cross-reactions based on the changes in agglutination rate in response to variations in dilution factor. Even though further experiments with a greater variety of data are necessary, this experiment demonstrates the potential of CAD. With the capability to identify cross-reaction, we can investigate the conditions of cross-reaction, which can be applied to elucidate *Leptospira*.

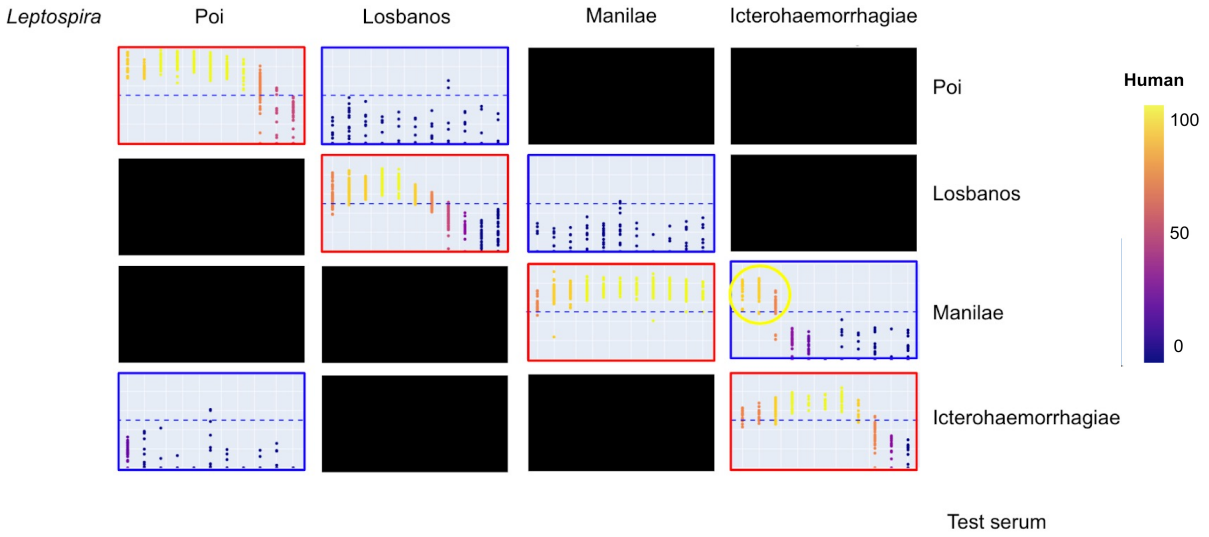

FIG 3 Transition of agglutination rate against dilution factor for different serovar-test serum pairs. In each plot, the horizontal and vertical axis represent the dilution factor in logarithmic and the estimated agglutination rate. Each plot is colored by its annotated agglutination rate.

Experiment on visualization

Here, we present the results of an experiment to visualize the proposed network's decision-making basis. UMAP projects original data into corresponding lower-dimensional spaces, placing similar data points nearby while ensuring that dissimilar data points are well-separated.

Visualization of the distribution of image features

Fig. 4, 5, and 6 shows the distribution of image feature vectors with different color annotations. These plots indicate that the feature extraction network extracts image features representing the abundance of bacteria in the images. Reference images are plotted closer together, while inspection images are more widely distributed. Most of the inspection images from unmatch cases are plotted closer to the reference images, while those from Ict-Man cases, which might cause a cross-reaction, are broadly distributed. Inspection images from matched cases are broadly distributed, and the distance from the reference images is proportional to the dilution factor.

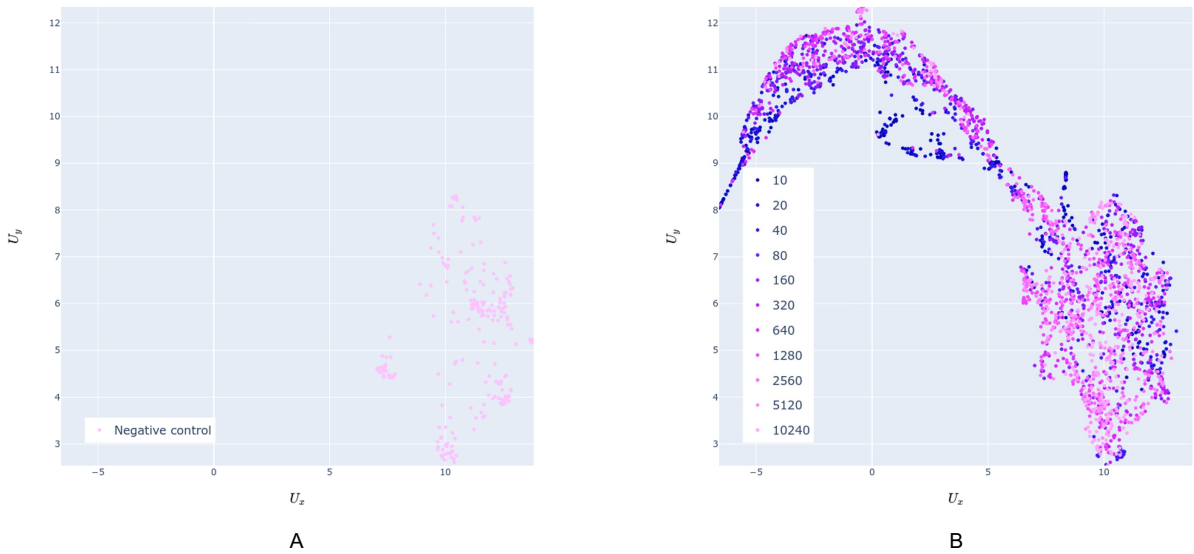

FIG 4 The distribution of all the image feature vectors. A: Reference images. B: Inspection images colored by dilution rate.

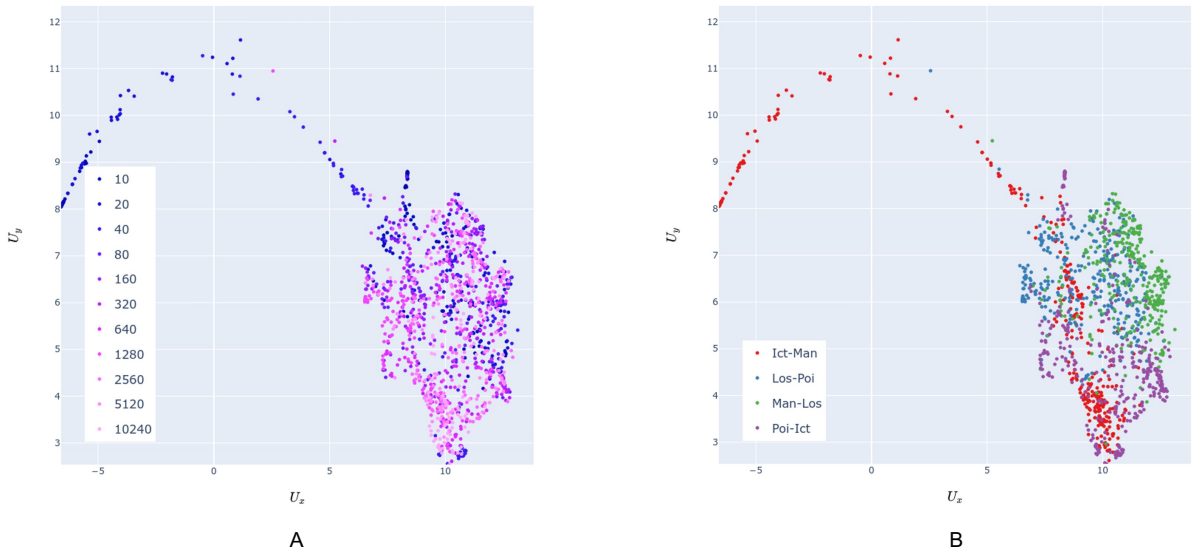

FIG 5 The distribution of image feature vectors of all the unmathced inspection images. A: Colored by dilution rates. B: Colored by test serum-serovar combinations.

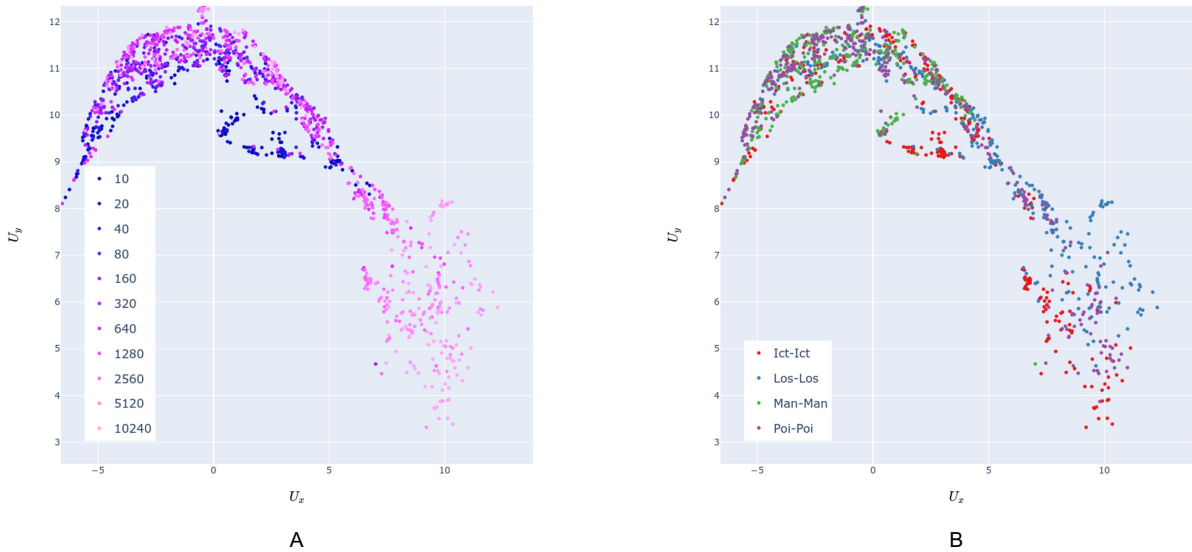

FIG 6 The distribution of image feature vectors of all the matched inspection images. A: Colored by dilution rates. B: Colored by test serum-serovar combinations.

Visualization of the distribution of image pair features

Fig. 7, 8, and 9 show the distribution of pair feature vectors with different color annotations. Considering the distribution trend, the proposed neural network extracts features representing the difference in the number of bacteria between image pairs. Image pairs from unmatched cases form two clusters: a larger cluster consisting of image pairs with low agglutination rates, and a smaller cluster of Ict-Man cases, which might cause cross-reaction. Image pairs of matched cases are distributed according to agglutination and dilution rates.

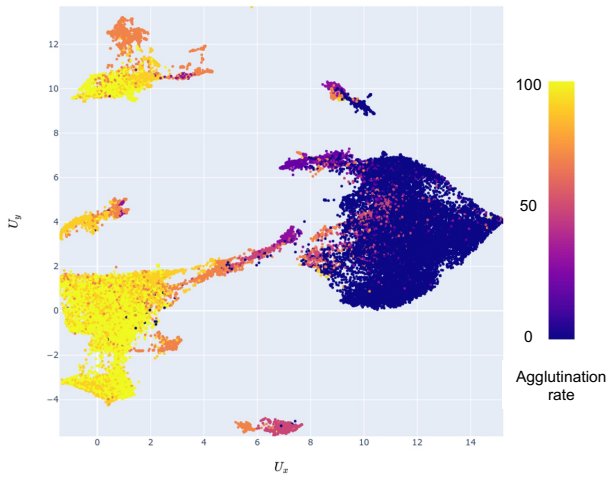

FIG 7 The distribution of all the pair feature vectors.

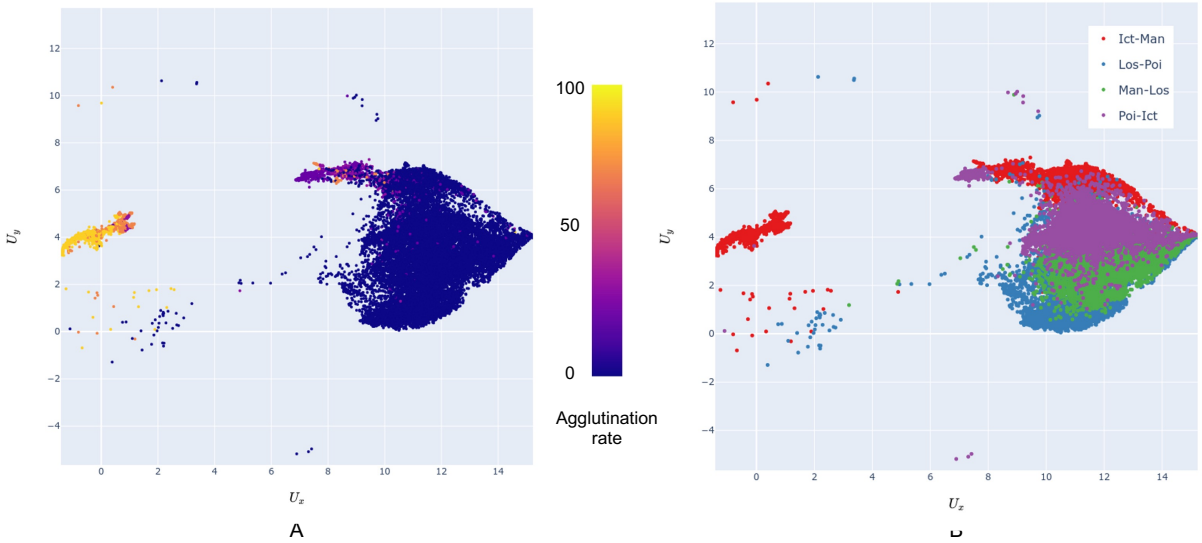

FIG 8 The distribution of all the unmatched pair feature vectors. A: Colored by agglutination rates. B: Colored by test serum-serovar combinations.

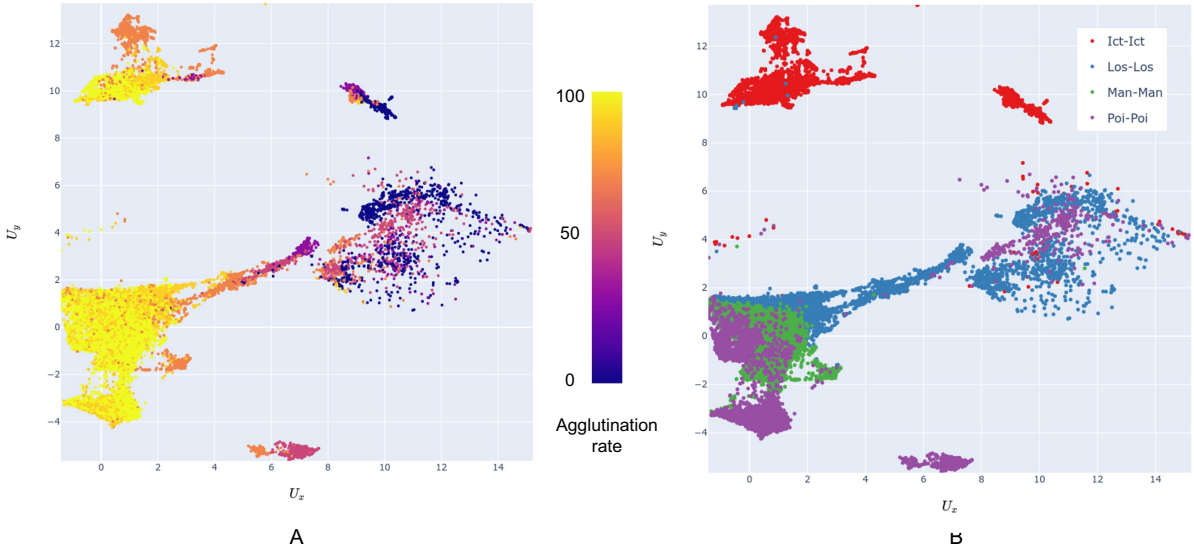

FIG 9 The distribution of all the matched pair feature vectors. A: Colored by agglutination rates. B: Colored by test serum-serovar combinations.

DISCUSSION

This study aims to replace the subjective evaluation of skilled laboratory technicians with deep learning techniques. Many studies have been published on implementing deep learning in CAD, and this paper is based on the same idea. As previously mentioned, this study has only been validated on a small dataset and thus faces many challenges. However, many of the experimental results of the proposed network are interesting and cannot be ignored as chance or unintentional bias. We have obtained valuable findings for efforts to introduce deep learning into MAT as a technology to understand and prevent the spread of neglected tropical diseases. Based on these results, another contribution of this study is the clarification of the challenges of resolving the difficulties of animal experiments and annotation.

The most important issue is evaluating with a wider variety of data. We used limited data for the experiment, and hence, we cannot distinguish whether the obtained result is due to individual differences or universal characteristics. To address this issue, we need to increase the number of samples per serovar and the number of serovars. With a sufficient amount of data, we can verify the possibility of determining negative, positive, and cross-reaction from the transition of agglutination rate against dilution factors.

It is also important to handle the dataset's imbalance. Since our dataset has an equal number of positive and negative cases, data with a 0% agglutination rate dominates. Hence, the proposed network pays less attention to data with medium and high agglutination rates. For this issue, we need to apply a training algorithm designed for such imbalanced data or collect more data with medium and high agglutination rates.

Mitigating annotation bias is also a significant concern. The current annotation task requires subjective evaluation by experienced laboratory technicians. To reduce the annotation cost, a more cost-effective annotation strategy, such as human-in-the-loop training [13], is favored.

CONCLUSION

Preventing the spread of a zoonotic disease necessitates speeding up and automating definitive diagnosis. In the case of leptospirosis, MAT is the standard test for definitive diagnosis. This paper aims to provide computer-aided diagnosis support for calculating agglutination rates based on the subjective assessment of MAT. We estimated the agglutination rate using the proposed deep learning network and visualized the decision-making process to avoid the "black box" problem.

We conducted experiments on agglutination rate estimation and visualization to serve as the basis for deep learning judgments. The agglutination rate estimation experiment indicates that deep learning is helpful for this task. We found that infection can be determined by examining the agglutination rate. The visualization experiment demonstrates that the proposed method fulfills its accountability. The results also showed a correlation with non-annotated serovars, suggesting that deep learning may have captured inherent characteristics of the leptospires, such as shape differences among serovars. Further experiments with a more diverse dataset are required to validate this indication.

Our results suggest that that deep learning is useful for agglutination rate estimation and that fulfilling the accountability helps medical experts understand how deep learning works. Our proposed network estimates the agglutination rate similarly to ordinary laboratory technicians. Although the estimation error is relatively higher for cases with higher agglutination rates, this type of objective quantification is still beneficial for the technicians. Furthermore, we found that the sequence of estimated agglutination rates visualizes the difference between positive reactions and cross-reactions. Additionally, using the dimensionality compression algorithm for visualization has addressed the black box problem in deep learning, fulfilling our accountability to non-specialists in computer science.

ACKNOWLEDGMENTS

This work was supported by Fukuoka University Research Grant (Subject number: 221041) and JSPS KAKENHI Grant Numbers JP18K16174 and JP21K16320 to RO.

DATA AVAILABILITY STATEMENT

Data will be made available on request.

FUNDING

This work was supported by Fukuoka University Research Grant (Subject number: 221041) and JSPS KAKENHI Grant Numbers JP18K16174 and JP21K16320 to RO.

CONFLICTS OF INTEREST

We hereby declare the following conflicts of interest in relation to this study. The author YO and RO are co-inventors of a patent 7370573 based on the technology developed in this research. This patent is registered in Japan. To ensure that this conflict of interest has not biased the research outcomes, an independent audit of the entire process of data collection, analysis, and interpretation was conducted. The other co-authors declare no direct financial conflicts of interest related to this research.

References

[1] **Faine S**. 1994. Leptospira and leptospirosis. CRC Press Inc.

[2] **Farrar J, Hotez PJ, Junghanss T, Kang G, Lalloo D, White NJ, Garcia PJ**. 2023. Manson's Tropical Diseases E-Book. Elsevier health sciences.

[3] **Musso D, La Scola B**. 2013. Laboratory diagnosis of leptospirosis: a challenge. *Journal of Microbiology, Immunology and Infection* 46 (4):245–252.

[4] **World Health Organization**. 2003. Human leptospirosis: guidance for diagnosis, surveillance and control. No WHO/CDS/CSR/EPH 2002.23, World Health Organization.

[5] **Oyamada Y, Ozuru R, Masuzawa T, Miyahara S, Nikaido Y, Obata F, Saito M, Villanueva SYAM, Fujii J**. 2021. A machine learning model of microscopic agglutination test for diagnosis of leptospirosis. *Plos one* 16 (11):e0259907.

[6] **LeCun Y, Bengio Y, Hinton G**. 2015. Deep learning. *nature* 521 (7553):436–444.

[7] **Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI**. 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42:60–88.

[8] **Becht E, McInnes L, Healy J, Dutertre CA, Kwok IW, Ng LG, Ginhoux F, Newell EW**. 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology* 37 (1):38–44.

[9] **Villanueva SY, Saito M, Tsutsumi Y, Segawa T, Baterna RA, Chakraborty A, Asoh T, Miyahara S, Yanagihara Y, Cavinta LL, et al.**. 2014. High virulence in hamsters of four dominant Leptospira serovars isolated from rats in the Philippines. *Microbiology* 160 (2):418–428.

[10] **Iandola F, Moskewicz M, Karayev S, Girshick R, Darrell T, Keutzer K**. 2014. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869* .

[11] **He K, Zhang X, Ren S, Sun J**. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *In* Proceedings of the IEEE international conference on computer vision p 1026–1034.

[12] **McInnes L, Healy J, Melville J**. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* .

[13] **Monarch RM**. 2021. Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI. Simon and Schuster.