UDC 004.895
doi:10.31799/1684-8853-2020-5-70-79

Articles

Comparative assessment of text-image fusion models for medical diagnostics

A. A. Lobantsev[a], *Programmer Engineer, orcid.org/0000-0002-8314-5103*
N. F. Gusarova[a], *PhD, Tech., Associate Professor, orcid.org/0000-0002-1361-6037, natfed@list.ru*
A. S. Vatian[a], *PhD, Tech., Associate Professor, orcid.org/0000-0002-5483-716X*
A. A. Kapitonov[b], *Post-Graduate Student, orcid.org/0000-0003-1378-1910*
A. A. Shalyto[a], *Dr. Sc., Tech., Professor, orcid.org/0000-0002-2723-2077*
[a]*ITMO University, 49, Kronverksky Pr., 197101, Saint-Petersburg, Russian Federation*
[b]*Belarus State Medical University, 83, Dzerzhinski Ave., 220116, Minsk, Republic of Belarus*

Introduction: *Information overload and complexity are characteristic of decision-making in medicine. In these conditions information fusion techniques are effective. For the diagnosis and treatment of pneumonia using x-ray images and accompaniyng free-text radiologists reports, it is promising to use text-image fusion.* ***Purpose:*** *Development of a method for fusing text with an image in the treatment and diagnosis of pneumonia using neural networks.* ***Methods:*** *We used MIMIC-CXR dataset, the SE-ResNeXt101-32x4d for images feature extracting and the Bio-ClinicalBERT model followed by ContextLSTM layer for text feature extracting. We compared five architectures in the conducted experiment: image classifier, report classifier and three scenarios of the fusion, namely late fusion, middle fusion and early fusion.* ***Results:*** *We got an absolute excess of metrics (ROC AUC = 0.9933, PR AUC = 0.9907) when using an early fusion classifier (ROC AUC = 0.9921, PR AUC = 0.9889) even over the idealized case of text classifier (that is, without taking into account possible errors of the radiologist). The network training time ranged from 20 minutes for late fusion to 9 hours and 45 minutes for early fusion. Based on Class Activation Map technique we graphically showed that the image feature extractor in the fused classification scenario still learns discriminative regions for pneumonia classification problem.* ***Discussion:*** *Fusing text and images increases the likelihood of correct image classification compared to only image classification. The proposed combined image-report classifier trained with the early-fusion method gives better performance than individual classifiers in the pneumonia classification problem. However, it is worth considering that better results cost the training time and required computation resources. Report-based training is much faster in training and less demanding for computation capacity.*

Keywords *— text-image fusion, medical diagnostics, late fusion, early fusion, middle fusion, x-ray image.*

For citation: Lobantsev A. A., Gusarova N. F., Vatian A. S., Kapitonov A. A., Shalyto A. A. Comparative assessment of text-image fusion models for medical diagnostics. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2020, no. 5, pp. 70–79. doi:10.31799/1684-8853-2020-5-70-79

Introduction

The richer the information flow is, the more potential opportunities it provides for the organization of effective control, but at the same time, the more difficult it is for the decision-maker to review it and bring it together into a single whole. Therefore, in conditions of information overload and complexity, information fusion techniques are increasingly used. At the substantive level, information fusion can be described as combining complementary information from different sources concerning the same object or scene to obtain complex information representation providing more effective control. In the literature, fused sources are often in different modalities, and the process itself is called multimodal fusion.

A number of definitions of information fusion are presented in the literature [1–6], emphasizing various aspects of this process. In the context of this article, it is vital to highlight the following aspects:

— information fusion aims to maximize the decision maker's or expert systems performance;

— sources of information for information fusion can be not only physical sensors but also social media and human intelligence reports (expert knowledge);

— information fusion is a hierarchical process and can occur at multiple abstraction levels (measurements, features, decisions).

Multimodal fusion is widely used in various practical tasks, such as web-search [7], image segmentation [8–10], image and video classification [11, 12], emotion recognition [13, 14], analysis of social media content [15], audio-visual speech enhancement [16], etc. With the development of high-tech diagnostic tools, multimodal fusion is becoming increasingly prominent in medicine. Here, such areas are actively developing as predicting the patient's health based on genomic, transcriptomic, and lifestyle information of one [17] as well as predicting the development of certain diseases [18, 19]. For instance, multimodal fusion in neuroimaging is actively developing these days [20–24]. It combines data from multiple imaging modalities, like positron emission tomography, computed tomography, and magnetic resonance imaging, to overcome the limitations of individual

modalities. In this case, an artificially combined image of the zone of interest in the brain is constructed as subject to further visualizing. In the clinical process, the attending physician acts as an analyst and interpreter of those images.

It is vitally important to make the right decision as fast as it is possible, but this is implementable only if there are adequate diagnostic tools. As we are trying to solve the classification of pneumonia presence problem we have to build on the diagnostic tools used in these cases. Chest x-ray (CXR) is mandatory for lung diseases, so it is hard to overemphasize its importance. Developing multimodal fusion-based technologies may improve such concerns of the CXR as the time gap from getting the image to taking the clinical decision and accuracy of the image interpretation. Both these possible improvements should be based on the experience of interpreting CXR's gained by professional radiologists because they are carriers of context and highly specialized knowledge.

Here it is impossible to diminish the significance of the interpretations performed by radiologists. They are carriers of context and highly specialized knowledge that the attending physician does not have. Their interpretation is presented in the form of a textual report. However, like any human opinion, these reports are subjective and error-prone [25, 26].

On the other hand, today, a large and growing role in interpreting medical images is assigned to machine learning-based tools. The results of their work can be, for instance, segmented images with highlighted malignant zones or other areas of interest, the manifestation of the alleged diagnosis, etc. However, the use of machine learning in medicine encounters well-known problems, including the small volumes of available datasets, which do not allow to form a full-fledged context for learning, and the uninterpretable conclusions (the so-called black-box problem). Hence, medical image processing results with machine learning-based tools cannot be regarded as objective and error-free.

In this regard, for the diagnosis and treatment of pneumonia presence, it is promising to use text-image fusion in order to provide the attending physician with information for making a decision in the most reliable, visible and at the same time interpretable (not "black-box") form. The article discusses the emerging challenges, and also proposes an approach to their solution based on the specific problem , namely the classification of pneumonia presence using x-ray images and accompaniyng free-text radiologists reports.

Background and related works

The analysis of reviews [6, 17, 20, 21, 27, 28] devoted to the problem of multimodal fusion allows us to highlight the main aspects that determine specific technological solutions for the fusion of images and texts in medicine:

— features the fusion is based on;
— IT architecture;
— and fusion level.

With regard to images, [21] divides features the fusion is based on into three groups: based on spatial domain, based on transform domain, based on deep learning. Features of the first group are the result of spatial (per-pixel) image segmentation, which can be performed according to various rules. For example, in [29, 30], the characteristics of intensity, hue and saturation are used for this. The selected features can be semantically labelled using anatomical brain atlases [31], special visual indexes [31] or ontologies [32]. The ontology describes the medical terms via a controlled vocabulary, where the conceptualizations of the domain knowledge are constructed as an OWL (Ontology Web Language) model.

Features of the second group are the result of transforming the source image to spatial-frequency domain which gives subband images with different scales and directions [33–35]. For example, in [35], sparse representation algorithm is used for merging low frequency subbands.

Features of the third group are formed directly by the neural network during the learning. As we approach the network's output layers, they reflect the image's structure with an increasing degree of generality, thereby constituting its multi-level abstract representation [36]. Nevertheless, as a rule, they stay semantically uninterpretable up to the last (decision) network level [7].

As for the text, the components of the text itself (words, terms, phrases, etc.) [37] or embeddings which transfer text information into a dense representation of the semantic space [38, 39] can be used as features. However, the texts of medical reports demonstrate the high syntactic and terminological complexity and the ambiguity in word usage. As shown in [40], embeddings perform significantly better in natural language processing tasks for such texts.

As the analysis of literature sources shows and as confirmed by the practice of radiologists [41], it is challenging to establish a formal correspondence between any fragments of medical images and fragments of medical reports describing them without involving semantic interpretation in both domains. Attempts to use external structures for this (such as visual indexes [31] or ontologies [32]) lead to significant losses in context, which in many cases decreases the benefits of multimodal fusion. Therefore, in recent publications, approaches related to the use of features that preserve contextual domain dependencies dominate [18, 19, 21–23], and

the deep learning methods are used as a technological base.

In tasks of semantic processing of medical texts, contextual word embeddings, primarily BERT [38], consisting of multiple layers of transformers which use self-attention mechanism, show the best results [40, 42, 43]. For example, for fusing text and speech in depression detection [19] features were extracted by BERT-CNN and VGG-16 CNN in combination with Gated Convolutional Neural Network (GCNN) followed by a LSTM layer. Additionally, [42] shows that BERT performs better than traditional word embedding methods in feature extraction tasks, and the BERT pre-trained on the clinical texts shows itself better than pre-trained on the general domain texts.

Thus, as the analysis shows, today, the CNN + BERT bundle is a popular architecture for joint processing of information from semantically loaded images and texts, a typical medicine case. However, the question of the fusion level remains open.

In a semantic sense, the information fusion can be performed at various levels of abstraction: measurements, features, decisions [1]. In [44], these levels correspond to various technological (architectural) solutions: recognition-based (also known as early fusion), decision-based (also known as late fusion), and hybrid multi-level.

In [45], feature map (DenseNet) and embeddings (BERT) are fused in an intermediate level using cross attention mechanism and afterward are supplied to fully connected layers. [12] compares the effectiveness of three fusion options: early fusion and late fusion combine information on the first convolutional layer and the first fully connected layer respectively, and slow fusion is a balanced mix between the two approaches such that higher layers get access to progressively more global information in both dimensions. The third option revealed some advantages and a strong dependence of efficiency on the context (the content of the dataset). In [46] and [8], the identical structures of late fusion with different weighting options are implemented, with opposite efficiency estimates obtained.

The work [9] explores several options for merging information of different modalities to highlight objects in an industrial landscape, including late sum fusion, late max fusion, late convolution fusion, and early fusion. Early fusion showed the best results; simultaneously, it was revealed that performance is highly problem dependent. Similar results were obtained in [10].

Authors of the work [23] implement a hybrid multi-level fusion based on a particle swarm search method to obtain optimal results. In [47], the authors propose a specialized module for intermediate fusion. The module operates between CNN streams and recalibrates channel-wise features in each modality. This module is generic and, in principle, suitable for any task, but, as the authors themselves note, the optimal locations and number of modules are different for each application.

Late fusion is rather popular in different applications [7, 18, 19, 37]. However, as noted in [47], this can be mainly due to resource reasons: the network for each unimodal stream can be designed and pre-trained independently for each modality. At the same time, late fusion can give rise to losing cross-modality information [27]. On the other hand, as the literature review reveals, the implementation of early and, even more so, multi-level fusion is a complex technological task and highly dependent on the subject area.

In conclusion of the review, it should be said that there are few works devoted directly to the fusion of texts and images for the diagnostics of particular diseases [18, 48] — this emphasizes the urgency of the problem.

Considering that the formation and training of deep learning models that implement multimodal merging for medical applications is a complex and resource-intensive process, the authors set themselves the task of experimentally assessing the effect of the fusion level on the effectiveness of the classification of pneumonia presence using x-ray images and accompaniyng free-text radiologists reports.

Method and materials

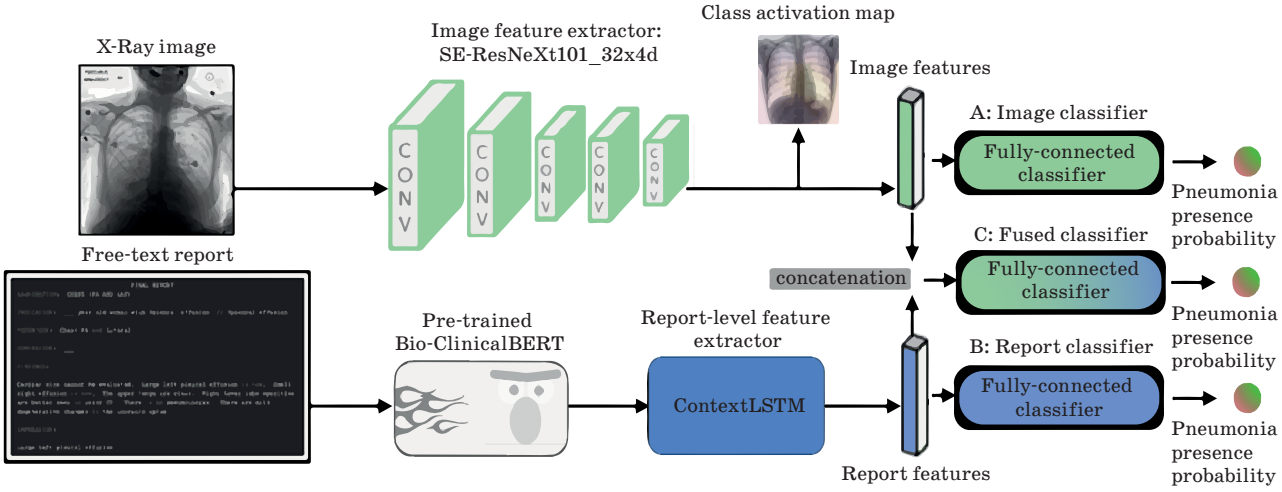
Datasets and data preprocessing

We used MIMIC-CXR [49–51] — the extensive publicly available chest radiographs dataset with free-text radiology reports. The dataset contains more than 300,000 images from over 60,000 patients. Labels were made with CheXpert labeler [50]. We used p10 and p11 data split folders.

We have performed the data preprocessing as follows. For images, pixel values were normalized to the range [0.0, 1.0]. Then, according to the DICOM field "Photometric Interpretation (0028, 0004)" images pixel values were inverted such that air in the image appears white (highest pixel value), while the patient's body appears black (lower pixel value). Then, the scikit-image library [52] was used to histogram equalization of the image to enhance contrast. Histogram equalization involves shifting pixel values towards 0.0 or 1.0 such that all pixel values have approximately equal frequency. All images were resized to 224 × 224 px size. For reports, we excluded all the text punctuation, and then used tokenizer from the pretrained BERT model (see next section).

Models and training procedure

We consider the model's architecture as a combination of feature extractors for image and text da-



■ ***Fig. 1.*** Training pipeline of the network

ta, and the fusion network appended to combine the extracted features. Training pipeline describing our training approach is shown in Fig. 1.

For images feature extracting we used the SE-ResNeXt101-32x4d [53] which is the ResNeXt101-32x4d model with added Squeeze-and-Excitation [54] module. In order to get word embeddings to feed the report feature extractor we used the pre-trained Bio-ClinicalBERT Model [55]. The Bio-ClinicalBERT model was trained on all notes from the MIMIC III database containing electronic health records from ICU patients at the Beth Israel Hospital in Boston [51].

The report feature extractor prepares report-level embeddings, which we train with the network we called the ContextLSTM layer. The ContextLSTM includes two-layer stacked LSTM network, and one fully connected layer. As input, ContextLSTM takes the sequence of word embeddings from the report. Then, each embedding vector from the sequence is stacked (concatenated) with the next N neighbours in the sequence, forming the context vectors for each entry of the sequence. We used context size $N = 2$.

We compare five architectures in the reported experiment.

Image classifier. To classify pneumonia using only x-ray images, on top of the image feature extractor, we placed a fully connected classification layer followed by the softmax activation.

Report classifier. To classify pneumonia using only reports, on top of the report feature extractor, we placed a fully connected classification layer prepended by the ReLU activation and followed by the softmax activation.

Fused classifier. As a fusion model we used the model consisting of the both feature extractors, followed by the single fully-connected classification layer. We consider three scenarios of the fusion:

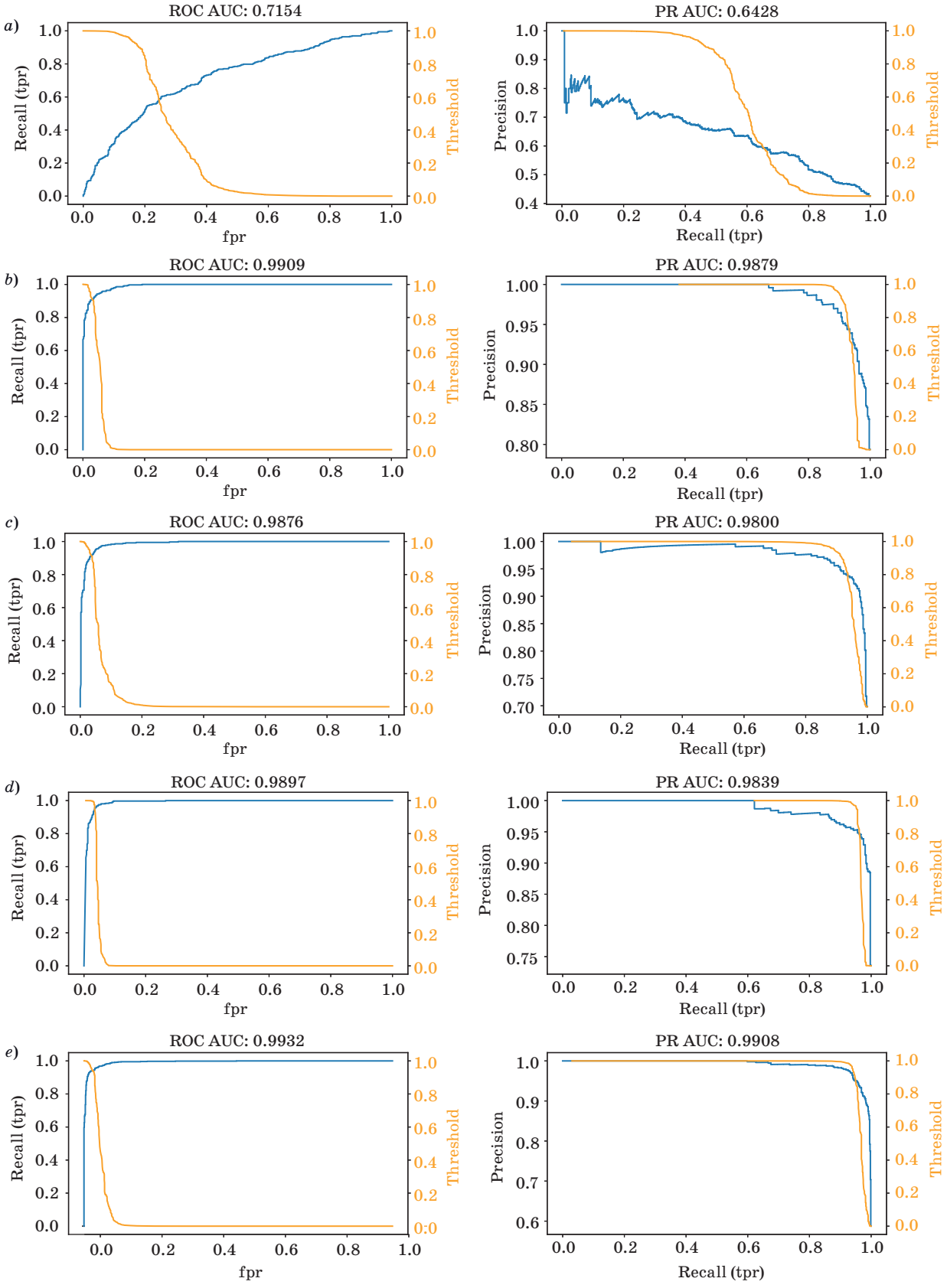
Late Fusion: fully-connected classifier is training on top of completely frozen pre-trained image and report feature extractors.

Middle Fusion: fully-connected classifier is training on top of the pre-trained image and text feature extractors with unfrozen last block of the SE-ResNeXt101-32x4d image feature extractor and the entire unfrozen ContextLSTM report feature extractor.

Early Fusion: the entire network is trained (except the BERT word embedding extractor) including SE-ResNeXt101-32x4d feature extractor, ContextLSTM report feature extractor, and the last fully-connected classification layer.

To compare the model quality we use the metrics: accuracy, model sensitivity and model specificity (at 0.5 probability threshold), area under ROC curve, area under precision-recall curve. Of these, the ROC AUC and PR AUC metrics — integral metrics for all possible decision thresholds — are key to conclusions. Metrics accuracy, specitifcity and sensitivity, typical for the medical literature, are auxiliary in this case. They are indicated at a typical threshold value of 0.5; the task of selecting the optimal threshold value was not posed here.

Training procedure we used looks as follows. First, we pre-trained each feature extractor as a separate classifier with the same task of classifying the presence of pneumonia. Then we trained late- and middle-fusion scenarios on the top of the pre-trained feature extractors. Finally, we trained the early fusion scenario network. All experiments were conducted with Adam optimizer (betas = 0.9, 0.999), learning rate was optimized with the Cosine Annealing scheduler [56] with the following hyperparameters of the scheduler: base_LR = 1e-8, T_0 = 50, T_mult = 2, eta_max = 1e-4, GAMMA = 0.1. We used batch size = 16, image



■ ***Fig. 2.*** Curves ROC (left) and PR (right) for: *a* — model of the image classifier; *b* — report classifier models; *c* — report classifier models for the Late-fusion classifier; *d* — Middle-fusion classifier; *e* — Early-fusion classifier

size = (224, 224). During training, images were augmented with random flips, shifts, scales, rotations, and small elastic transforms. NVIDIA RTX 2080Ti was used for training.

Results and discussion

We present the obtained metrics values for all types of classifiers in the Table, using the following abbreviations: ROC AUC — area under the ROC curve, PR AUC — area under precision-recall curve. These curves for each type of classifier is illustrated by Fig. 2, *a*–*e*. The last column in the Table shows training time comparison.

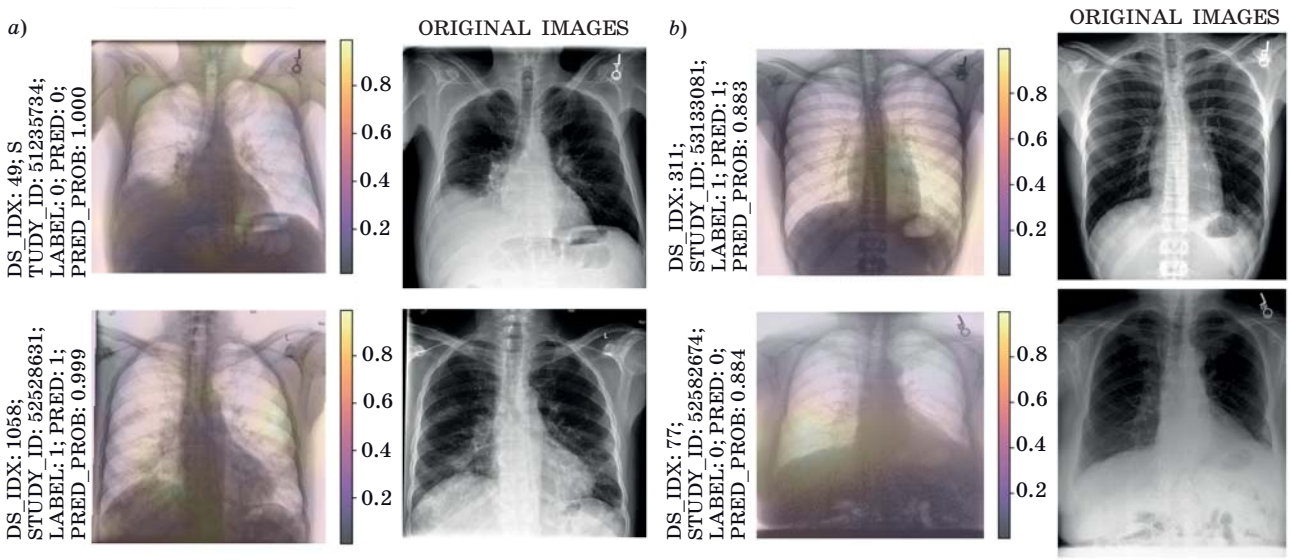
Figure 3 shows examples of Class Activation Maps [57] for results of image classifier and image feature extractor in early-fusion model indicating the most distinctive areas used by each classifier to determine the category to which the image belongs: 0 — no pneumonia, 1 — pneumonia. Areas with higher activation values have more impact on the network decision. Color graduations correspond to the level of activation of the image features on the last convolution layers. Lighter areas are more critical for the network to gain the decision than the darker ones. Here we can see that the image feature extractor in the fused classification scenario still learns discriminative regions for pneumonia classification problem.

Analysis of the results obtained allows us to draw the following statements.

Though the values of all metrics for the image classifier (see Table) are relatively poor compared to the report classifier, they entirely correspond to the current world level [58–60]. Note that during training the report classifier, only obviously correct text reports were used; in reality, it is necessary to consider the human factor associated with errors and fatigue of radiologists, which reduces the received metrics. Under these conditions, fusing text and images increases the likelihood of correct image classification. Besides, Table shows that we got an absolute excess of metrics when using an early fu-

■ Metrics values for different classifiers

Classifier type	Accuracy	Sensitivity	Specificity	ROC AUC	PR AUC	Training time
Image classifier	0.6823	0.5718	0.7637	0.7195	0.6438	6 h 14 min
Report classifier	0.9590	0.9696	0.9511	0.9921	0.9889	20 min
Late-fusion classifier	0.9484	0.9558	0.9430	0.9876	0.9800	15 min
Middle-fusion classifier	0.9613	0.9669	0.9572	0.9897	0.9838	29 min
Early-fusion classifier	0.9579	0.9684	0.9504	0.9933	0.9907	9 h 45 min



■ ***Fig. 3.*** Class Activation Map for different classifiers: *a* — early fusion classifier; *b* — image classifier

sion classifier even over the idealized case of text classifier.

Comparing Figures 3, *a* and *b* shows that in the case of the image classifier, the network learns only from those areas of the image that are directly affected by pneumonia. Simultaneously, in the case of the fusion classifier, the network takes into account not only these areas but also the surrounding context, which is more consistent with the approach of a radiologist to the medical images classification. Besides, as shown in Table, the report classifier learns much faster than all other classifiers. However, Fig. 3, *b* indicates that introducing a text classifier into the fusion pipeline does not lead to training only on the textual data, i.e., image feature extractor in the fused network model still learns correct semantic image areas for classification.

Comparing the results of the Table for three types of fusion shows an exchange ratio between the classification and training efficiency metrics: the higher the desired classification efficiency metrics, the more computing resources are required to train the network. This circumstance must be taken into account when choosing the proposed models in a real clinical process.

Conclusion

We have presented that combining x-ray images and accompanying free-text radiologists reports in neural network model training improves the quality of the model's decision in the classification task.

In the experiment, we have compared five training scenarios for the pneumonia classification task. Two individual classifiers based on each modality and three fused classifiers differ in training methods: late, middle, and early fusion. The proposed combined image-report classifier trained with the early-fusion method gives better performance than individual classifiers in the pneumonia classification problem. However, it is worth considering that better results cost the training time and required computation resources. Report-based training is much faster in training and less demanding for computation capacity.

Acknowledgment

This work was financially supported by Russian Science Foundation, Grant 19-19-00696.

References

1. **Dimou I., Zervakis M., Lowe D., and Tsiknakis M.** *Computational Methods and Tools for Decision Support in Biomedicine: An Overview of Algorithmic Challenges*. In: *Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications*. T. P. Exarchos, A. Papadopoulos and D. I. Fotiadis (eds.). 2009. Pp. 1–17.
2. **Yu-Jin Zhang.** *Image Fusion Techniques with Multiple-Sensors*. In: *Encyclopedia of Information Science and Technology*. Third Ed. 2015. Pp. 5926–5936.
3. **Solaiman B., Bosse E.** *Possibility Theory for the Design of Information Fusion Systems*. Series: Information Fusion and Data Science. 2019. 278 p.
4. **Rogova G. L.** *Information Quality in Fusion-Driven Human-Machine Environments*. In: *Information Quality in Information Fusion and Decision Making*. Series: Information Fusion and Data Science. E. Bosse, G. L. Rogova (eds.). 2019. Pp. 3–29.
5. **Calderero F., Marqués F.** *Image Analysis and Understanding Based on Information Theoretical Region Merging Approaches for Segmentation and Cooperative Fusion*. In: *Handbook of Research on Computational Intelligence for Engineering, Science, and Business*. S. Bhattacharyya, and P. Dutta (eds.). 2013. Pp. 75–121.
6. **Meng T., Jing X., Yan Z., Pedryczc W.** A survey on machine learning for data fusion. *Information Fusion*, May 2020, vol. 57, pp. 115–129.
7. **Wang D., Mao K., Ng G.-W.** Convolutional neural networks and multimodal fusion for text aided image classification. *20th International Conference on Information Fusion*, 2017, July 10–13, 2017, Xi'an, China, pp. 1–7.
8. **Perdana C. R. A., Nugroho H. A., Ardiyanto I.** Comparison of text-image fusion models for high school diploma certificate classification. *Communications in Science and Technology*, 2020, vol. 5(1), pp. 5–9.
9. **Lawin F. J., Danelljan M., Tosteberg P., Bhat G., Khan F. S., and Felsberg M.** Deep projective 3d semantic segmentation. *International Conference on Computer Analysis of Images and Patterns*, 2017. arXiv:1705.03428v1 [cs.CV] 9 May 2017.
10. **Tetreault J.** Deep *Multimodal Fusion Networks for Semantic Segmentation*. 2017. All Theses. 2756. Available at: https://tigerprints.clemson.edu/all_theses/2756 (accessed 10 September 2020).
11. **Gallo I., Calefati A., Nawaz S., Janjua M. K.** *Image and Encoded Text Fusion for Multi-Modal Classification*. arXiv:1810.02001 [cs.CV] 3 Oct. 2018.
12. **Karpathy A., Toderici G., Shetty S., Leung T., Sukthankar R., and Li Fei-Fei.** Large-scale video classification with convolutional neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus*, OH, 2014, pp. 1725–1732, doi:10.1109/CVPR.2014.223
13. **Jiang Y., Li W., Hossain M. S., Chen M., Al-Hammadi M.** A snapshot research and implementation of multimodal information fusion for data-driven emo-

tion recognition. *Information Fusion*, Jan. 2020, vol. 53, pp. 209–221.
14. **Gravina R., Li Q.** Emotion-relevant activity recognition based on smart cushion using multi-sensor fusion. *Information Fusion*, Aug. 2019, vol. 48, pp. 1–10.
15. **Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Jiebo Luo.** Multimodal fusion with recurrent neural networks for rumor detection on microblogs. *Proceedings of the 25th ACM International Conference on Multimedia (MM'17)*, Oct. 2017, pp. 795–816. https://doi.org/10.1145/3123266.3123454
16. **Afouras T., Chung J.S., and Zisserman A.** *The conversation: Deep audio-visual speech enhancement*. arXiv preprint arXiv:1804.04121, 2018.
17. **Zitnik M., Nguyen F., Wang B., Leskovec J., Goldenberg A., Hoffman M. M.** Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, Oct. 2019, vol. 50, pp 71–91.
18. **Qiang He, Xin Li, D. W. Nathan Kim, Xun Ji, Xuejun Gu, Xin Zhen, Linghong Zhou.** Feasibility study of a multi-criteria decision-making based hierarchical model for multi-modality feature and multi-classifier fusion: Applications in medical prognosis prediction. *Information Fusion*, Mar. 2020, vol. 55, pp. 207–219.
19. **Makiuchi M. R., Warnita T., Uto K., Shinoda K.** Multimodal fusion of BERT-CNN and gated CNN representations for depression detection. *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC '19)*, Oct. 2019, pp. 55–63. https://doi.org/10.1145/3347320.3357694
20. **Zhang Y.-D., Dong Z., Wang S.-H., Yu X., Yao X., Zhou Q., Hu H., Li M., Jiménez-Mesa C., Ramirez J., Martinez F. J., Gorriz J. M.** Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation. *Information Fusion*, 2020, vol. 64, pp. 149–187.
21. **Huang B., Yang F., Yin M., Mo X., and Zhong C.** A review of multimodal medical image fusion techniques. *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 8279342. https://doi.org/10.1155/2020/8279342
22. **Fan F., Huang Y., Wang L., Xiong X., Jiang Z., Zhang Z., and Zhan J.** *A semantic-based medical image fusion*. arXiv:1906.00225v2 [eess.IV] 11 Dec. 2019.
23. **Huang C., Tian G., Lan Y., Peng Y., Ng E. Y. K., Hao Y., Cheng Y., and Che W.** A new Pulse Coupled Neural Network (PCNN) for brain medical image fusion empowered by shuffled frog leaping algorithm. *Front. Neurosci.*, 20 Mar. 2019. https://doi.org/10.3389/fnins.2019.00210
24. **Huang Z., Lin J., Xu L., Wang H., Bai T., Pang Y., and Meen T.-H.** Fusion high-resolution network for diagnosing ChestX-ray images. *Electronics*, 2020, vol. 9, iss. 1, p. 190. doi:10.3390/electronics9010190
25. **Razzak M. I., Naz S., and Zaib A.** *Deep Learning for Medical Image Processing: Overview, Challenges and Future*. In: *Classification in BioApps. Lecture Notes in Computational Vision and Biomechanics*. N. Dey, A. Ashour, S. Borra (eds). Springer, Cham, 2017. Vol. 26. Pp. 323–350.
26. **Shen D., Wu G., and Suk H.-I.** Deep learning in medical image analysis. *Annu. Rev Biomed Eng.*, 2017, vol. 19, pp. 221–248.
27. **Gao J., Li P., Chen Z., and Zhang J.** A survey on deep learning for multimodal data fusion. *Neural Computation*, May 2020, vol. 32, iss. 5, pp. 829–864.
28. **Smirnov A., Levashova T.** Knowledge fusion patterns: A survey. *Information Fusion*, Dec. 2019, vol. 52, pp. 31–40.
29. **Chen C.-I.** Fusion of PET and MR brain images based on HIS and Log-Gabor transforms. *IEEE Sensors Journal*, 2017, vol. 17, no. 21, pp. 6995–7010.
30. **Cabezas M., Oliver A., Lladó X., Freixenet J., Cuadra M. B.** A review of atlas-based segmentation for magnetic resonance brain images. *Comput. Methods Programs Biomed.*, 2011, vol. 104, iss. 3, E158–E177.
31. **Racoceanu D., Lacoste C., Teodorescu R., Vuillemenot N.** A semantic fusion approach between medical images and reports using UMLS. *Proceedings of Third Asia Information Retrieval Symposium, (AIRS 2006) "Information Retrieval Technology"*, Singapore, October 16–18, 2006, Springer, Berlin, Heidelberg, pp. 460–475.
32. **Teodorescu R.-O., Cernazanu-Glavan C., Cretu V.-I., and Racoceanu D.** The use of the medical ontology for a semantic-based fusion system in biomedical informatics application to Alzheimer disease. *4th International Conference on Intelligent Computer Communication and Processing*, Cluj-Napoca, 2008, pp. 265–268. doi:10.1109/ICCP.2008.4648383
33. **Yin M., Liu X., Liu Y., and Chen X.** Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain. *IEEE Transactions on Instrumentation and Measurement*, 2018, no. 99, pp. 1–16.
34. **Mendhe M., Ladhake S. A., Ghate U. S.** An application of Shearlet transform for medical image fusion. *International Journal of Engineering Research & Technology (IJERT)*, May 2017, vol. 6, iss. 05, pp. 833–837.
35. **Shabanzade F., and Ghassemian H.** Multimodal image fusion via sparse representation and clustering-based dictionary learning algorithm in nonsubsampled contourlet domain. *8th International Symposium on Telecommunications (IST)*, Tehran, Iran, Sept. 2016, pp. 472–477.
36. **LeCun Y., Bengio, Y., & Hinton G. E.** Deep learning. *Nature*, 2015, vol. 521, iss. 7553, pp. 436–444.
37. **Ma L., Lu Z., Shang L., & Li H.** Multimodal convolutional neural networks for matching image and sentence. *Proceedings of 2015 IEEE International Conference on Computer Vision*, Washington, DC, IEEE Computer Society, 2015, pp. 2623–2631.
38. **Devlin J., Chang M.-W., Lee K., and Toutanova K.** *BERT: Pre-training of deep bidirectional transform-*

ers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
39. **Mikolov T., Chen K., Corrado G., and Dean J.** *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781, 2013.
40. **Arora S., May A., Zhang J., Ré C.** Contextual embeddings: when are they worth it? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020, pp. 2650–2663.
41. **Boitsov V., Vatian A., Egorov N., Klochkov A., Lobantsev A., Markova E., Gusarova N., Shalyto A., Zubanenko A., Soldatov R., Niyogi R.** Software tools for manual segmentation of tomography images supporting radiologist's personal context. *25th Conference of Open Innovations Association (FRUCT)*, Helsinki, Finland, 2019, pp. 64–76. doi:10.23919/FRUCT48121.2019.8981541
42. **Si Y., Wang J., Xu H., Roberts K. E.** Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, July 2019, vol. 26, iss. 11, pp. 1297–1304. doi:10.1093/jamia/ocz096
43. **Khattak F. K., Jeblee S., Pou-Prom C., Abdalla M., Meaney C., Rudzicz F.** A survey of word embeddings for clinical text. *Journal of Biomedical Informatics*, Dec. 2019, vol. 4, 100057.
44. **D'Ulizia A.** *Exploring Multimodal Input Fusion Strategies*. In: *Handbook of Research on Multimodal Human Computer Interaction and Pervasive Services: Evolutionary Techniques for Improving Accessibility*. P. Grifoni (ed). IGI Publ., 2009. Pp. 34–57.
45. **Jaimes A.** *Multi-Modal Fusion AI for Real-time Event Detection*. Febr. 18, 2020. Available at: https://www.dataminr.com/blog/multi-modal-fusion-ai-for-real-time-event-detection (accessed 5 August 2020).
46. **Li G., and Li N.** Customs classification for cross-border e-commerce based on text-image adaptive convolutional neural network. *Electron. Commer. Res.*, 2019, vol. 19, iss. 4, pp. 799–800.
47. **Joze H. R. V., Shaban A., Iuzzolino M. L., Koishida K.** *MMTM: Multimodal Transfer Module for CNN Fusion*. arXiv:1911.08670v2 [cs.CV] 30 Mar. 2020.
48. **Vatian A., Gusarova N., Dobrenko N., Klochkov A., Nigmatullin N., Lobantsev A., and Shalyto A.** Fusing of medical images and reports in diagnostics of brain diseases. *Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence (PRAI '19)*, Aug. 2019, pp. 102–108. https://doi.org/10.1145/3357777.3357793
49. **Johnson A. E. W., Pollard T. J., Berkowitz S., Greenbaum N. R., Lungren M. P., Deng C.-Y., Mark R. G., Horng S.** *MIMIC-CXR: A large publicly available database of labeled chest radiographs*. arXiv:1901.07042 [cs.CV] 14 Nov. 2019.
50. *MIMIC-CXR-JPG — chest radiographs with structured labels 2.0.0*. Availible at: https://physionet.org/content/mimic-cxr-jpg/2.0.0/mimic-cxr-2.0.0-chexpert.csv.gz, last access 20.08.2020 (accessed 5 August 2020).
51. **Johnson A. E. W., Pollard T. J.,, Shen L., Lehman L. H., Feng M., Ghassemi M., Moody B., Szolovits P., Celi L. A.** & **Mark R. G.** MIMIC-III, a freely accessible critical care database. *Sci Data*, 2016, vol. 3, Article number 160035. https://doi.org/10.1038/sdata.2016.35
52. **van der Walt S., Schönberger J. L., Nunez-Iglesias J., Boulogne F., Warner J. D., Yager N., Gouillart E., Yu T.** Scikit-image: Image processing in Python. *Peer J*, 2014, 2:e453. https://doi.org/10.7717/peerj.453
53. **Xie S., Girshick R., Dollar P., Tu Z., He K.** *Aggregated Residual Transformations for Deep Neural Networks*. arXiv:1611.05431v2 [cs.CV] 11 Apr. 2017.
54. **Hu J., Shen L., Albanie S., Sun G., Wu E.** *Squeeze-and-Excitation Networks*. arXiv:1709.01507v4 [cs.CV] 16 May 2019.
55. **Alsentzer E., Murphy J. R., Boag W., Weng W.-H., Jin D., Naumann T., McDermott M. B. A.** Publicly available clinical BERT embeddings. *Clinical Natural Language Processing (ClinicalNLP) Workshop at NAACL*, 2019. https://arxiv.org/abs/1904.03323
56. **Loshchilov I., Hutter F.** SGDR: Stochastic Gradient Descent with Warm Restarts. *ICLR 2017 Conference Paper*. arXiv:1608.03983 [cs.LG].
57. **Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A.** Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1–10.
58. **Wang X., Peng Y., Lu L., Lu Z., Summers R. M.** Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1–16.
59. **Rajpurkar P., Irvin J., Zhu K., Yang B., Mehta H., Duan T., Ding D., Bagul A., Langlotz C., Shpanskaya K., Lungren M. P., Ng A. Y.** *Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning*. arXiv preprint arXiv:1711.05225, 2017.
60. **Khan W., Nazar Z., and Luqman A.** *Intelligent pneumonia identification from chest x-rays: A systematic literature review*. medRxiv. 2020.

УДК 004.895
doi:10.31799/1684-8853-2020-5-70-79

Сравнительная оценка моделей слияния текста и изображения для медицинской диагностики

А. А. Лобанцев[а], инженер-программист, orcid.org/0000-0002-8314-5103
Н. Ф. Гусарова[а], канд. техн. наук, доцент, orcid.org/0000-0002-1361-6037, natfed@list.ru
А. С. Ватьян[а], канд. техн. наук, доцент, orcid.org/0000-0002-5483-716X
А. А. Капитонов[б], аспирант, orcid.org/0000-0003-1378-1910
А. А. Шалыто[а], доктор техн. наук, профессор, orcid.org/0000-0002-2723-2077
[а]Университет ИТМО, Кронверкский пр., 49, Санкт-Петербург, 197101, РФ
[б]Белорусский государственный медицинский университет, Дзержинского пр., 83, 220116, Минск, Белоруссия

Введение: в медицине при принятии решений характерны информационная перегрузка и сложность. В этих условиях эффективны методы слияния информации. Для диагностики и лечения пневмонии с использованием рентгеновских снимков и их текстовых описаний, выполняемых радиологами, перспективно использовать слияние текста с изображением. **Цель:** разработка моделей слияния изображения и текста при диагностике пневмонии с помощью нейронных сетей. **Методы:** использовался датасет 33 MIMIC-CXR; для обработки изображений использована сеть SE-ResNeXt101-32x4d; для обработки текста использована модель Bio-ClinicalBERT в сочетании со слоем ContextLSTM. Проведено экспериментальное сравнение пяти архитектур нейронной сети: классификатор изображений, классификатор текстов и три классификатора на основе слияния, а именно позднего, раннего и промежуточного слияния. **Результаты:** при использовании классификатора раннего слияния получено абсолютное превышение показателей (ROC AUC = 0,9933, PR AUC = 0,9907) даже по сравнению с идеализированным (т. е. без учета возможных ошибок радиолога) случаем текстового классификатора (ROC AUC = 0,9921, PR AUC = 0,9889). Время обучения сети варьировалось от 20 минут для позднего слияния до 9 часов 45 минут для раннего слияния. С использованием карты активации классов наглядно показано, что во всех классификаторах на основе слияния действительно выделяются наиболее характерные для классификации пневмонии области изображения. **Обсуждение:** слияние текста и изображений увеличивает вероятность правильной классификации изображений по сравнению с классификацией только изображений. Показано, что в задаче классификации пневмонии классификатор изображений и текстов, обученный с помощью метода раннего слияния, дает лучшую производительность, чем классификаторы изображений и текстов по отдельности. Однако стоит учесть, что лучшие результаты требуют затрат времени на обучение и вычислительных ресурсов. Обучение на основе текстовых отчетов проходит намного быстрее и требует меньших вычислительных ресурсов.

Ключевые слова — слияние текста и изображения, медицинская диагностика, позднее слияние, раннее слияние, промежуточное слияние, рентгеновское изображение.

Для цитирования: Lobantsev A. A., Gusarova N. F., Vatian A. S., Kapitonov A. A., Shalyto A. A. Comparative assessment of text-image fusion models for medical diagnostics. *Информационно-управляющие системы*, 2020, № 5, с. 70–79. doi:10.31799/1684-8853-2020-5-70-79

For citation: Lobantsev A. A., Gusarova N. F., Vatian A. S., Kapitonov A. A., Shalyto A. A. Comparative assessment of text-image fusion models for medical diagnostics. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2020, no. 5, pp. 70–79. doi:10.31799/1684-8853-2020-5-70-79

ПАМЯТКА ДЛЯ АВТОРОВ

Поступающие в редакцию статьи проходят обязательное рецензирование.

При наличии положительной рецензии статья рассматривается редакционной коллегией. Принятая в печать статья направляется автору для согласования редакторских правок. После согласования автор представляет в редакцию окончательный вариант текста статьи.

Процедуры согласования текста статьи могут осуществляться как непосредственно в редакции, так и по e-mail (ius.spb@gmail.com).

При отклонении статьи редакция представляет автору мотивированное заключение и рецензию, при необходимости доработать статью — рецензию.

Редакция журнала напоминает, что ответственность за достоверность и точность рекламных материалов несут рекламодатели.