ETop3PPE: EPOCh's Top-Three Prediction Probability Ensemble Method for Deep Learning Classification Models

Javokhir Musaev, Abdulaziz Anorboev, Huyen Trang Phan, and Dosam Hwang(✉)

Yeungnam University, Daegu, Republic of Korea
dosamhwang@gmail.com

Abstract. The rapid growth in the field of deep learning (DL) has increased research interests in this area, including computer vision (CV), for which high-quality products have been produced. Many fields, including medicine, the automobile industry, finance, education, and the military have used state-of-art CV results. Any change in CV affects human life through the abovementioned sector. Considering the importance of the development in CV, we proposed Epoch's Top-Three Prediction Probability Ensemble (Etop3PPE) method for DL classification problems. Our method focuses on the use of lost knowledge during training and an optimal way to increase the accuracy of the model. Each epoch during training represents a different prediction space, which is the key to the development of the proposed method. We used the top-three prediction probabilities of each image for the classification task from different epochs of training and ensembled them into the best model prediction probabilities. Applying our method, we partially solved the ensemble error problem in the models and increased the accuracy of our model from 89.32% to 90.91%. We added 32.8% lost knowledge from the prediction spaces of the training epochs that were used during the ensembling. We used the Cifar10dataset to evaluate our method. In addition, we compared the results of our method with those of the classic ensemble (ensembling all prediction probabilities into the best model). The result was surprising, our method overcome by 16% in case of adding lost knowledge of different epochs.

Keywords: Prediction space · Top3 maximum prediction probabilities · Ensemble

1 Introduction

Deep learning (DL) has recently started a new era with the introduction of neural networks (NNs), and next-generation DL models and methods have been developed. As the quality of DL products continues to increase, numerous new topics are being studied. The implementation of DL in various fields such as medicine, IoT, the military and automobile industries, finance, and many other areas has increased the amount of funding spent on research in this field. In addition, the number of tasks in DL-based CV is growing

N. T. Nguyen et al. (Eds.): ACIIDS 2022, LNAI 13757, pp. 222–233, 2022.
https://doi.org/10.1007/978-3-031-21743-2_18

continuously. Particularly during the spread of Covid 19, detection has become one of the main tools used to check the temperature and mask compliance of individuals. Other implementations, from medical diagnosis to heavy industry, are increasing significantly in number. Any change in DL adds an extra opportunity for the future development of the field.

To fill in the gap of lost knowledge occurring during training, we conducted research on DL model ensemble methods. Ensemble learning is an extremely powerful DL method. Its effect on the results has been studied by numerous researchers [1]–[4]. Various ensemble models and methods contributing to the development of ensemble learning in DL have been proposed. Despite a large number of studies, gaps requiring a solution exist, and improvements in CV have been proposed. Our initial studies showed that the true prediction scope of the worse models differs from that of the best models. In our study, we attempted to manipulate the prediction scope of the training epochs. We chose the DL classification task to study and improve the use of ensemble methods. Our experiment results show that each epoch can truly predict different images better than the other one, even when their accuracy is quite small. We attempted to find an optimal method to use this knowledge in our research. In our previous study, we used the maximum epoch prediction probabilities and achieved a higher accuracy than that achieved through classic training. In this study, we developed our previous study by applying the top-three maximum prediction probabilities to the ensemble model. We used the VGG50 model pretrained with ImageNet to train the Cifar10 dataset. We replaced the last layer of VGG50 with a dense layer, including 10 nodes that are equal to the number of classes in Cifar10. Next, we resized the Cifar10 dataset into $224 \times 224 \times 3$ sized images and trained the model for 20 epochs using callbacks to save the best epoch from among these 20 epochs. After achieving an accuracy of 86.08% for the first 20 epochs, we trained 20 more epochs, with an accuracy of 89.32%. The second model was chosen as the base model. The following step was to check the prediction spaces of the first model that saved after the first 20 epochs. Its 4.85% true prediction rate differed from the true predictions of the second model. We were motivated to use this insight to increase the accuracy by applying the true prediction space of this second model. We used the prediction probabilities of the first model with an accuracy of 86.08% and selected the top-three or three maximum prediction probabilities for each image in the test set and ensemble them into the corresponding position within the prediction probabilities of the second model with an accuracy of 89.32%. As a result, we increased the accuracy of the model from 89.32% to 90.91%.

Our study consists of the following sections. Section I introduces the proposed method and provides general information regarding ensemble learning and the content of this study. Section II provides information regarding related studies and problems found in ensemble learning. Section III provides solutions to the problems described in Section II and provides a detailed explanation of the proposed methodology. Section IV describes the experiments and results of our research, including information regarding the dataset, base method, training setup, evaluation metrics, experiment results, and discussions. Finally, Section V provides some concluding remarks regarding this research and areas for future study.

2 Related Works

Many researchers have studied ensemble learning and their applications in various fields. The combination of data representation and model ensembles [4] has been the focus of numerous researchers. The bias-variance tradeoff and cross-validation ensembles were among the forms of DL ensembles that were studied in [5]. Other forms of DL ensembles include varying the training data, models, and combinations. If we look at applications throughout various sectors, we can find numerous applications of ensemble learning in credit risk assessment [6], oil price forecasting [7], multi-step forecasting for big data time series, [8] and an ensemble learning model for Covid-19 based on CT images [9]. In [6], the authors studied a credit risk assessment and proposed a multistage neural network ensemble for risk assessment using a bagging sampling method for generating training data. In addition, different models were used in [6] to train the dataset, and the results were scaled into unit intervals followed by fusion. For oil price prediction [7], the authors used stacked denoising autoencoders to model the nonlinear and complex relationships of oil prices with its features. Another implementation [8] of ensemble learning conducted on the forecasting of big data time series used decision trees, gradient boosted trees, and a random forest to develop an ensemble model. Weights for ensemble members were computed using a weighted least squares method. The next successful study on ensemble learning focused on Covid-19 detection using 2933 lung CT images obtained from different sources. The authors initialized the model parameters using transfer learning and three pretrained DL models: AlexNet, GoogleNet, and ResNet. These models were used to extract features from all images and the final dense layer with the softmax function used for classification. The final accuracy is higher than that of the component classifiers of the ensemble. Effective approach to prevent asthma is to control it using data from asthma patients. [10] studied 90 asthma patients during 9 months and collected data of the patients from specialized hospital for pulmonary diseases in Tehran. Authors [10] proposed new ensemble learning algorithm with combining physicians' knowledge in the form of a rule-based classifier and supervised learning algorithms to detect asthma control level in a multivariate dataset with multiclass response variable. The model outcome resulting from the balancing operations and feature selection on data yielded the accuracy of 91.66%. The next implementation [11] of ensemble learning was dedicated to improve medical image segmentation. [11] proposed new methods to improve the segmentation probability estimation without losing performance in a real-world scenario that has only one ambiguous annotation per image. Authors marginalize the estimated segmentation probability maps of networks that are encouraged to under-segment or over-segment with the varying Tversky loss without penalizing balanced segmentation. In addition, study proposed a unified hypernetwork ensemble method to alleviate the computational burden of training multiple networks. Proposed approaches successfully estimated the segmentation probability maps that reflected the underlying structures and provided the intuitive control on segmentation for the challenging 3D medical image segmentation. Following research [12] studied an analysis on ensemble learning optimized medical image classification with deep convolutional neural networks. In this work, a reproducible medical image classification pipeline was proposed to analyze the performance impact of augmenting, stacking, and bagging methods. The pipeline includes state-of-the-art pre-processing and image augmentation

methods and nine deep convolution neural network architectures. Four different popular medical image datasets were used to evaluate the research. It was applied on four popular medical imaging datasets with varying complexity. The results were evaluated with an F-1 score which showed that stacking, augmentation, and bagging increased the results up to 13%, 4 and 10%, respectively. Another research [13] studied performance analysis of hyperparameter optimization methods for ensemble learning with small and medium-sized medical datasets. For this task, the study analyzed Grid search, Random search, and Bayesian optimizations for an ensemble classifier. One more problem that studied in ensemble learning is medical diagnosis [14] with imbalanced data. The proposed method includes data pre-processing, training base classifier and final ensemble. In the data pre-processing step, they introduced the extension of Synthetic Minority Oversampling Technique (SMOTE) by integrating it with cross-validated committees filter (CVCF) technique. It allowed them to synthesize the minority sample and thereby balance the input instances by filtering the noisy examples. In the classification phase, they introduced ensemble support vector machine (ESVM) classification technique followed by the weighted majority voting strategy. Also, they proposed simulated annealing genetic algorithm (SAGA) to optimize the weight vector and thereby enhance the overall classification performance. The proposed ensemble learning method was tested on nine imbalanced medical datasets and achieved better results than other state-of-the-art classification models. The next ensemble learning study [15] proposes a disease prediction model (DPM) to provide an early prediction for type 2 diabetes and hypertension based on individual's risk factors data. DPM consists of isolation forest (iForest) based outlier detection method to remove outlier data, synthetic minority oversampling technique tomek link (SMOTETomek) to balance data distribution, and ensemble approach to predict the diseases. Four datasets were used to build the model and extract the most significant risks factors. The authors claims that the proposed DPM achieved highest accuracy when compared to other models and previous studies. Following research [16] studied magnetic resonance (MR) images and engines that effect the quality of the images. To solve one of the challenging problems in medical images [16] proposed an ensemble learning and deep learning framework that improves MR image resolution. Authors utilized five commonly used super-resolution algorithms and achieved enlarged image datasets with complementary priors. Subsequently, GAN is trained to generate super-resolution MR images. At the final step, another GAN is used for ensemble learning that synergizes the outputs of GANs into the final MR super-resolution images. Results of the study showed that achievements of the ensemble outperformed any single GAN's results. [17]–[23] proved that ensemble learning adds a significant improvement to the models and extends prediction spaces of the models.

After our literature review, we found a gap in which the epoch prediction probabilities and the effect of their prediction probabilities were not thoroughly studied. In this research, we learned the effect of the number of top prediction probabilities on the ensemble model.

3 Proposed Method

To address the gap found during the literature review and learn more insight from images using a DL ensemble, we proposed Epoch's Top Three Prediction Probability Ensemble (ETop3PPE) method for DL classification problems. When we studied ensemble learning and its forms, we found that not all insights of the models were optimally applied to the ensemble model. Varying the data, model, or their combinations yields better results than the main component of the ensemble. Despite this, there is still an opportunity to add extra knowledge to the final ensemble model using the knowledge of different epochs. The motivation for developing the proposed method was the true prediction spaces of different models. In our previous research, we studied the maximum prediction probabilities of the epoch and their effects on the ensemble models. In this research, we studied the effect of the number of top prediction probabilities on the ensemble models. When a model is trained with a dataset for the classification task, we achieve the classification probabilities for each image in the dataset. In case of 50000x32x32x3 sized dataset that has 10 classes and 50000 images with 32x32x3 sizes, prediction probability size equals to 50000x10, 10 classification probability for each images. We studied the true predictions of the epochs during training and determined that each epoch found different images better than the other epochs. For instance, if epoch 10 can truly predict images from the 1000^{th} position to the 45000^{th} position in the dataset, another epoch can be found from the 500^{th} position to the 43000^{th} position of the images. This shows that when we ensemble these two epochs, a certain number of images from the 500^{th} position to the 1000^{th} position can be accurately predicted in addition to the positions from the 1000^{th} to 45000^{th} positions.

Figure 1 illustrates ETop3PPE for deep learning classification models and includes the following steps:

1. In the initial step, we uploaded the data and resized it to $224 \times 224 \times 3$ and rescaled each pixel by dividing it by 255.
2. The pre-processed dataset was fed to the pretrained ResNet50 model with ImageNet dataset.
3. The model was trained for 20 epochs and best epoch was saved when considering the validation accuracy.
4. The model was trained for 20 more epochs, and for this interval of training, we saved the best model to evaluate its validation accuracy.
5. A high accuracy epoch was chosen as the main model, and a lower accuracy epoch was chosen as the secondary model.
6. The top-three prediction probabilities of the secondary model of each image are added to the corresponding prediction probabilities of the main model.
7. The maximum prediction probabilities of the main model were selected for each image for classification.

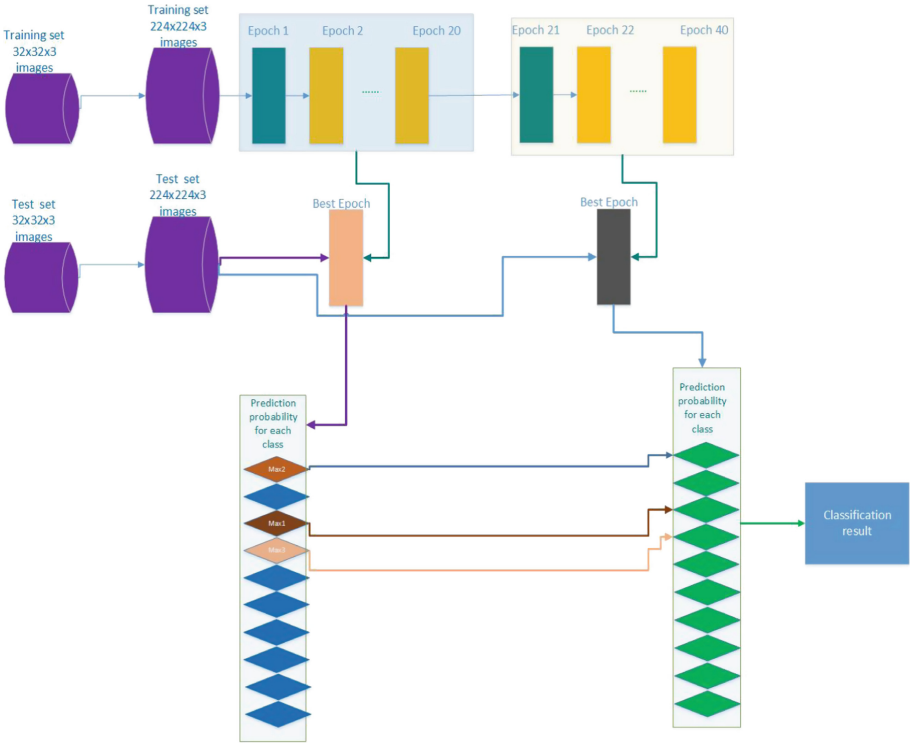

Fig. 1. Epoch's Top-Three Prediction Probability Ensemble Method

In this research, we used the Cifar10 dataset; hence, for the final layers of the ResNet50 model, we applied a dense layer with 10 nodes. The advantage of this model is that it can be used with other DL ensemble models and only adds more knowledge to the ensemble. We present the detailed practical effects of the method in the next section.

4 Experiments and Results

In this section, we provide comprehensive information regarding the experiments and their results. In addition, this section includes the test results of the model and dataset used. Here, we provide a clearer view of our method through experiments. We used the Cifar10 dataset to evaluate the proposed method.

4.1 Dataset

We used one of the popular datasets from the image classification field with a sufficient number of images and reliable labeled data, i.e., the Cifar10[1] collected by Krizhevsky, Nair, and Hinton, which is popular in classification tasks. In addition, the size of the

[1] https://www.cs.toronto.edu/~kriz/cifar.html.

dataset was advantageous for training. A clearer description of the classification is provided in Table 1. We changed the image size in the dataset from 32 × 32 × 3 to 224 × 224 × 3. To avoid bias from pre-processing, we used the minimum pre-processing tools. The images were normalized to 255. Cifar10 was the best choice in our research for training ResNet50. Because there is no limitation in using the dataset, this will help in further developing the method in future studies.

Table 1. Cifar10 dataset description.

Dataset name	Cifar10
Total number of images	60000
Size of images	32x32x3
Train set	50000
Test set	10000
Size of dataset	163 mb (python version)
Class names	"airplane", "automobile", "bird", "cat", "deer", "dog", "frog", "horse", "ship", "truck"

4.2 Training Setup

We used Python 3.9 and Python 2.1.0 of the TensorFlow framework in our training. The experiments were conducted using a 12 GB NVidia Titan-XP GPU with CUDA 10.2 on a computer with an Intel Core-i9 11th generation CPU and 64 GB of RAM. In our training, we initialized the weights with pretrained weights from ImageNet. In addition, we used a sparse categorical loss function for our training and chose 20 epochs for the first step, followed by an additional 20 epochs. We trained the model during different intervals and chose the best models for representing more knowledge.

4.3 Evaluation Metrics

In our training, we focused on accuracy as a main metric and used a unique true prediction (UTP) to explain the success of the method on an ensemble, which is the ratio of true predictions to the total number of cases used to evaluate the model. Equation (1) shows the calculation of the accuracy achieved.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

TP- true predicted positive results

TN-true predicted negative results

FP-false predicted positive results

FN- false predicted negative results

$$UTP(X, Y) = X - X \subset Y \tag{2}$$

UTP - Unique True Prediction

X - Prediction Scope of a model X

Y - Prediction Scope of a model Y

The next evaluation metric is UTP, which identifies the percentage of unique predictions for each model with respect to another prediction. In Eq. (2), UTP(X, Y) finds the UTP of model X with respect to model Y. These metrics explain why our proposed model achieved better results than the main model, where we trained only the main dataset. The indices of the true predicted images are different in each model, despite having the same accuracy. This leads the ensemble to achieve better results.

4.4 Experiment Results and Discussions

In this part of our study, we introduced a detailed explanation of the results and the reasons for achieving these results. Moreover, we evaluated our method using the accuracy and UTP. We used accuracy metrics because our main focus was on the effect of the epochs on the final prediction. We used the UTP metric to explain why better results were achieved.

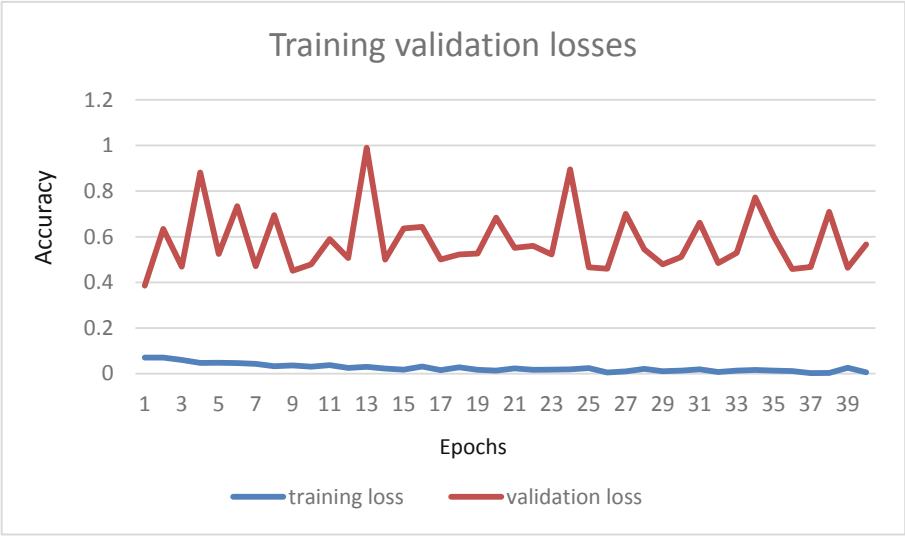

Fig. 2. Training and validation losses

Figure 2 shows the training and validation losses in our training and illustrates losses that include high bias as it reached a training loss extremely close to zero; however,

the accuracies of the validation were still higher than 0.4. The same trend is presented in Fig. 3, where the training and validation accuracies of the models are presented. Although the training accuracy reached 100%, the validation accuracy still did not reach 90%. As in many models, there is a generalization problem in that, although the training data are learned extremely well, not all features of the validation data can be extracted as training data.

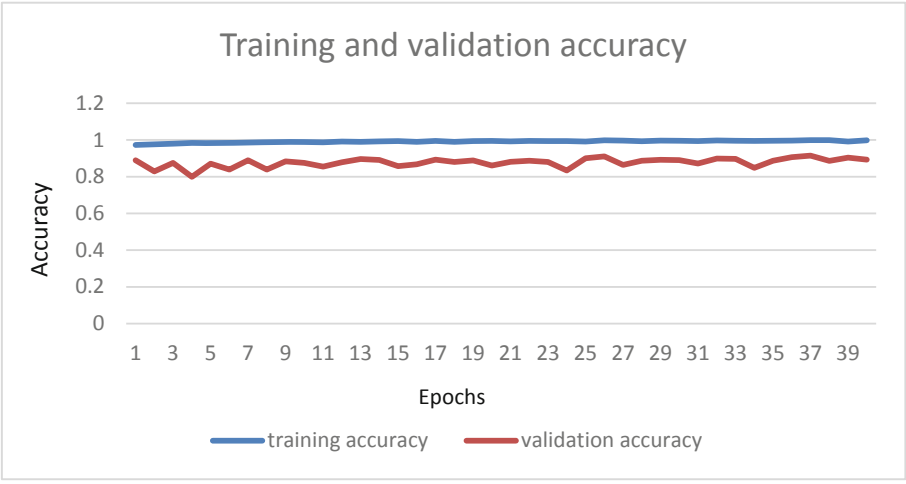

Fig. 3. Training and validation accuracies

To overcome this problem, we proposed increasing the accuracy of the model using additional knowledge from the epochs. Table 2 presents the UTP of the secondary model to the main model and the UTP of the main model to the ensemble. Analyzing this table, we can conclude that 4.85% of the knowledge was available for use.

Table 2. UTP (Secondary model, Main model)

Models	UTP
Secondary model to Main model	0.0485
Main model to (Main model + secondary model)	0.0175

We used only 32% of existing knowledge, and the rest of the knowledge was available as long as the prediction probabilities of the main model included incorrect predictions that affected the ensemble. Each component of the ensemble has both positive effects and unwanted side effects on the prediction space of the ensemble. After ensembling the models, we lost 1.75% of the true predictions from the main model, but still achieved better results than using only the prediction probabilities of the main model. Hence, when we analyzed Table 3, the accuracy of the main model was lower than that of the

ensemble model, which was created by ensembling prediction probabilities of the main model and the top-three prediction probabilities of the secondary model.

Table 3. Accuracies

Models	Accuracy
Secondary model	86.08
Main model	89.32
Secondary model + Main model	90.83
Proposed model	**90.91**

After applying our method, we increased the accuracy of the model to 90.91%. When we ensembled all prediction probabilities of the secondary model into the prediction probabilities of the main model, we obtained an accuracy of 90.83%. Our method achieved the expected results after training with Cifar10. Moreover, this method can be used simultaneously with other ensemble models.

5 Conclusions

In this study, we used a unique true prediction space as the main tool to find a gap and tried to fill it by applying the ETop3PPE method. When we used our method with the ResNet50 pretrained model and Cifar10 dataset, we achieved better results than when only using the ensemble component and adding all of the probabilities of the secondary model into the appropriate prediction probabilities of the main model. We explained our method results with an enlarged prediction space for the ensemble model. As a result, we were able to increase the accuracy of the model on the Cifar10 dataset from 89.32% to 90.91%. In addition, we used 32.8% of the extra knowledge of the secondary model. There is still a huge area of research remaining to be conducted on the epoch knowledge.

In the future work we plan to use ontology for building a knowledge base for meta information about the CV objects [24–26]. This base should be useful for processing the images.

References

1. Alqurashi, T., Wang, W.: Clustering ensemble method. Int. J. Mach. Learn. Cybern. **10**, 1–18 (2018). https://doi.org/10.1007/s13042-017-0756-7
2. Abbasi, S.-O., Nejatian, S., Parvin, H., Rezaie, V., Bagherifard, K.: Clustering ensemble selection considering quality and diversity. Artif. Intell. Rev. **52**(2), 1311–1340 (2018). https://doi.org/10.1007/s10462-018-9642-2
3. Vega-Pons, S., Ruiz-Shulcloper, J.: A survey of clustering ensemble algorithms. Int. J. Pattern Recogn. Artif. Intell. **25**(3), 337–372 (2011). https://doi.org/10.1142/S0218001411008683
4. Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q.: A survey on ensemble learning. Front. Comp. Sci. **14**(2), 241–258 (2019). https://doi.org/10.1007/s11704-019-8208-z

5. Krogh, A.: Neural network ensembles, cross validation, and active learning
6. Yu, L., Wang, S., Lai, K.K.: Credit risk assessment with a multistage neural network ensemble learning approach. Expert Syst. Appl. **34**(2), 1434–1444 (2008). https://doi.org/10.1016/j.eswa.2007.01.009
7. Zhao, Y., Li, J., Yu, L.: A deep learning ensemble approach for crude oil price forecasting. Energ. Econ. **66**, 9–16 (2017). https://doi.org/10.1016/j.eneco.2017.05.023
8. Galicia, A., Talavera-Llames, R., Troncoso, A., Koprinska, I., Martínez-Álvarez, F.: Multi-step forecasting for big data time series based on ensemble learning. Knowl.-Based Syst. **163**, 830–841 (2019). https://doi.org/10.1016/j.knosys.2018.10.009
9. Zhou, T., Lu, H., Yang, Z., Qiu, S., Huo, B., Dong, Y.: The ensemble deep learning model for novel COVID-19 on CT images. Appl. Soft Comput. **98** (2021). https://doi.org/10.1016/j.asoc.2020.106885
10. Khasha, R., Sepehri, M.M., Mahdaviani, S.A.: An ensemble learning method for asthma control level detection with leveraging medical knowledge-based classifier and supervised learning. J. Med. Syst. **43**(6), 1–15 (2019). https://doi.org/10.1007/s10916-019-1259-8
11. Hong, S., et al.: Hypernet-ensemble learning of segmentation probability for medical image segmentation with ambiguous labels (2021). Available: http://arxiv.org/abs/2112.06693
12. Müller, D., Soto-Rey, I., Kramer, F.: An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks
13. Kadam, V.J., Jadhav, S.M.: Performance analysis of hyperparameter optimization methods for ensemble learning with small and medium sized medical datasets. J. Discrete Math. Sci. Crypt. **23**(1), 115–123 (2020). https://doi.org/10.1080/09720529.2020.1721871
14. Liu, N., Li, X., Qi, E., Xu, M., Li, L., Gao, B.: A novel ensemble learning paradigm for medical diagnosis with imbalanced data. IEEE Access **8**, 171263–171280 (2020). https://doi.org/10.1109/ACCESS.2020.3014362
15. Fitriyani, N.L., Syafrudin, M., Alfian, G., Rhee, J.: Development of disease prediction model based on ensemble learning approach for diabetes and hypertension. IEEE Access **7**, 144777–144789 (2019). https://doi.org/10.1109/ACCESS.2019.2945129
16. Lyu, Q., Shan, H., Wang, G.: MRI super-resolution with ensemble learning and complementary priors. IEEE Trans. Comput. Imaging **6**, 615–624 (2020). https://doi.org/10.1109/tci.2020.2964201
17. Lahmiri, S., Bekiros, S., Giakoumelou, A., Bezzina, F.: Performance assessment of ensemble learning systems in financial data classification. Intell. Syst. Acc. Financ. Manag. **27**(1), 3–9 (2020). https://doi.org/10.1002/isaf.1460
18. Ni, J., Zhang, L., Tao, J., Yang, X.: Prediction of stocks with high transfer based on ensemble learning. J. Phys.: Conf. Ser., 1651, 1 (2020). https://doi.org/10.1088/1742-6596/1651/1/012124ssss
19. Gaikwad, D.P., Thool, R.C.: Intrusion detection system using bagging ensemble method of machine learning. In: Proceedings - 1st International Conference on Computing, Communication, Control and Automation, ICCUBEA 2015, pp. 291–295 (2015). https://doi.org/10.1109/ICCUBEA.2015.61
20. Rehman Javed, A., Jalil, Z., Atif Moqurrab, S., Abbas, S., Liu, X.: Ensemble adaboost classifier for accurate and fast detection of botnet attacks in connected vehicles. Trans. Emerg. Telecommun. Technol. (2020). https://doi.org/10.1002/ett.4088
21. Bynagari, N.B.: Anti-Money Laundering Recognition Through The Gradient Boosting Classifier
22. Pham, B.T., et al.: Ensemble modeling of landslide susceptibility using random subspace learner and different decision tree classifiers. Geocarto Int. (2020). https://doi.org/10.1080/10106049.2020.1737972

23. Soares, R.G.F., Chen, H., Yao, X.: A cluster-based semisupervised ensemble for multiclass classification. IEEE Trans. Emerging Top. Comput. Intell. **1**(6), 408–420 (2017). https://doi.org/10.1109/TETCI.2017.2743219
24. Pietranik, M., Nguyen, N.T.: A multi-atrribute based framework for ontology aligning. Neurocomputing **146**, 276–290 (2014)
25. Duong, T.H., Nguyen, N.T., Jo, G.S.: A method for integration of wordnet-based ontologies using distance measures. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008. LNCS (LNAI), vol. 5177, pp. 210–219. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85563-7_31
26. Nguyen, N.T.: Conflicts of ontologies – classification and consensus-based methods for resolving. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 267–274. Springer, Heidelberg (2006). https://doi.org/10.1007/11893004_34