Multimodal Machine Learning for 2D to 3D Mapping in Biomedical Atlases

B. Almogadwy, N. K. Taylor, and A. Burger
Computer Science Department, Heriot-Watt University, Edinburgh, UK
Email: {ba28, N.K.Taylor, A.G.Burger}@hw.ac.uk

***Abstract*—2D to 3D image registration has a vital role in medical imaging and remains a significant challenge. It primarily relates to the use and analysis of multimodal data. We address the issue by developing a multimodal machine learning algorithm that predicts the position of a 2D slice in a 3D biomedical atlas dataset based on textual annotation and image data. Our algorithm first separately analyses images and textual information using base models and then combines the outputs of the base models using a Meta-learner model. To evaluate learning models, we have built a custom accuracy function. We tested different variants of Convolutional Neural Network architectures and different transfer learning techniques to build an optimal image base model for image analysis. To analyze textual information, we used tree-based ensemble models, namely, Random Forest and XGBoost algorithms. We applied the grid search to find optimal hyperparameters for tree-based methods. We have found that the XGBoost model showed the best performance in combining predictions from different base models. Testing the developed method showed 99.55% accuracy in predicting 2D slice position in a 3D atlas model.**

***Index Terms*—image registration, multimodal data, EMAP atlas. CNN, deep learning**

I. Introduction

Imaging plays an essential role in modern biomedical sciences and forms the basis for much of current research and clinical diagnostic work. Accumulation of imaging data leads to the need for the integration of data collected from different sources. To meet this demand, a number of atlases of biomedical imaging were developed [1]-[4], which can not only provide detailed anatomical and histological information about the studied object but also provide a framework into which molecular information, such as gene and protein expression, can be mapped [5]. This paper suggests an algorithm that combines the advantages of different machine learning methods using stack generalization techniques. This algorithm utilizes different data modalities and machine learning models and integrates their predictions into a single pipeline. We show that the resulting algorithm performs better than each machine learning algorithm used on its own. Adding and extracting information from atlases of biomedical imaging leads to 2D/3D image registration tasks. Image registration is the process of relating features between images by aligning them. For example, the task of integrating new data into a biomedical atlas is a typical 2D to the 3D image registration problem. Images created in experiments are mostly 2D images, while modern biomedical atlases are mostly 3D models. To transfer the data related to the 2D image (e.g., spatial transcriptomics data) to the 3D atlas, the position of the new image in the 3D model must be determined, i.e., the matching cut (section) through the 3D model identified. Such mapping involves the analysis of multi-modal data, including image and textual anatomical information.

To address the above-described tasks, a variety of methods were developed [6], [7], such as image processing-based matching [8], ontology-based matching [6], deep learning [9], and others [10]. Some of these methods are based on analysis of image data, e.g., image processing-based matching methods, while others are focused on the analysis of textual description, such as ontology-based matching. Deep learning methods based on Convolutional Neural Networks (CNNs) are universal prediction tools working with different types of data modalities. However, deep learning methods require large amounts of data to achieve good prediction accuracy, which is often a serious limitation. Although deep learning models and decision trees appear to be theoretically equivalent tools [11], in practice, tree-based methods, such as XGBoost [12] or Random Forests [13], offer a good compromise between the size of the learning dataset and prediction accuracy [14]. Thus, different methods might be optimal depending on the available amount and number of data modalities [15].

The proposed solution explains that stack generalization combining deep learning and tree-based methods is a powerful image registration tool for medium-sized datasets.

II. Experiment and Methods

In our current work, we focus on the image registration of the data stored in the EMOUSE atlas (EMAP) [16], [5]. EMAP is a digital atlas of mouse embryo development that represents 3D models of embryos at various stages of development and gives spatial context to *in-situ* gene expression data experimentally obtained for mouse embryos (www.emouseatlas.org).

A. Source and Generation of Dataset

The EMAP 3D model [16], [5] was used to source data in our work. We extracted multiple 2D images (slices)

Manuscript received November 24, 2021; revised April 12, 2022.

©2022 Journal of Image and Graphics
doi: 10.18178/joig.10.2.64-69

from the EMAP 3D model and, along with the image data, also used the EMAP model to obtain a large textual dataset containing anatomical descriptions. Each slice in the EMAP 3D atlas is described using four values: the pitch, the yaw, the roll angles, and the distance of the sectioning plane (Fig. 1). To simplify the task, we only considered pitch and distance parameters in the 3D atlas in the present work. The pitch refers to the angle shown in Fig. 1. The value of the pitch parameter changes from 0 to 180 degrees. The distance is the vertical position of the slice, which ranges between -258 and 257. The values of yaw and roll parameters were set to zero.

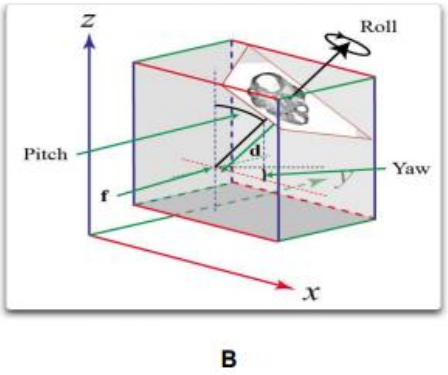

Figure 1. Parameters describing the position of the 2D slice in the 3D EMAP mode.

Considering that increments for both distance and pitch values are 1, we can obtain $(257 - (-258)) * 180 = 92700$ 2D slices. Before applying machine learning algorithms, we have pre-processed these data. These pre-processed data comprised our initial image dataset, where each image can be represented by the pair of distance (d) and pitch (p) values. Our dataset had the two following limitations. Firstly, there are only 20 anatomical structures for the current image data set, e.g., brain, heart, etc. This posed limitations for the textual description of the details of mouse anatomy. Secondly, we had only one image for a unique set of distance and pitch values. Although we used different data augmentation techniques, we were still strongly lacking the data to use all the potential of machine learning techniques.

B. Pre-processing Data

We have applied the following pre-processing steps to normalize the images, simplify the learning process and perform data augmentation.

1) Preprocessing the image data

- Conversion of the images to grayscale to avoid the obstacles of computational complexity, capacity, and memory storage.
- Normalization of the images such that the pixel values lie within the range of 0 to 1 for a unified representation.
- Removal of unwanted details from the images by cropping and resizing them to 128 *128 pixels. This resulted in a much smaller model with no compromise in performance.
- Since this is a supervised machine learning problem (where the dataset is labelled with output values), each image was labelled with its corresponding distance and pitch (d, p) values.
- The data consisted of only one image for each set of (d, p) values. So, data augmentation techniques were applied to increase the size of our dataset [17]. The set of procedures performed was as follows:
  i). Image rotation to 36 different angles.
  ii). Skewing to add noise to the image (skewness is a feature that computes the measure or lack of Symmetry in the image). After applying the skewing step, the images were left, right, top, and bottom-justified or skewed.
  iii). Slightly shifting to add some variation between images of the same class.

The purpose behind the augmentation of the images was to expand the dataset (training set) by creating new samples [18].

2) An example of image data

The image in Fig. 2 is a sample taken from the dataset to elaborate the image part of our dataset and how it was preprocessed for a better understanding. The labels for this image are (241,1) (distance, pitch), respectively. Combining the anatomical labels, this image contains these anatomical structures '1st branchial arch, 2nd branchial arch, olfactory pit, ear, neural tube, and hindbrain'.

Further, this image is augmented to create almost a hundred other samples with a little variation using the steps mentioned above.

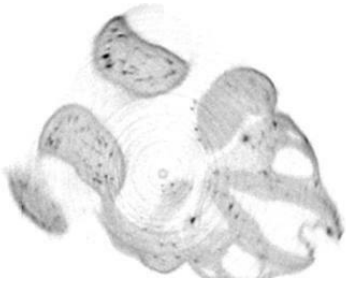

Figure 2. Sample image from EMOUSE atlas image data.

3) Preprocessing the text data

- Removing all empty lists: Some of the slices had no labelled anatomical structure to avoid performance compromise.
- Excluding: Mislabeled data is primarily misleading and causes deviation in training the model. Therefore it was removed from the dataset.
- Bag of Words: A binary vector of length 20 representing a list of unique anatomical names was generated. This binary vector represented any anatomical description for a given image. '1' was assigned where the anatomical description contained the feature word and '0' if the word did not appear in the document.

4) Example of the anatomical data as a binary vector

There are 20 unique anatomical labels observed in the textual dataset; a binary vector was created with all 20 of them represented as dictionary vectors as follows:

['1st branchial arch', '1st branchial arch maxillary component', '2nd branchial arch', 'diencephalon',

'diencephalon floorplate', 'ear', 'eye', 'forelimb bud', 'forelimb bud apical ectodermal ridge', 'heart', 'hindbrain', 'liver', 'midbrain', 'neural tube', 'olfactory pit', 'rathke', 'somite', 'spouch', 'tail', 'telencephalon']

The vectors for individual records of the textual data are formed as binary (ones and zeros) vectors of 20 components, where 1 represents the presence of an anatomical label and zero represents its absence. For better understanding, consider the list of anatomical labels discussed in section 3.2.1.1. The binary vector for the example would look something like this:

[1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0]

C. Accuracy Metric

We found that the most popular metrics commonly used to evaluate the accuracy of regression models, the Mean Square Error (MSE) and Root Mean Square Error (RMSE), do not allow for an accepted error threshold for our task. Due to the similarity between our dataset's classes, it was difficult to predict the exact distance correctly for a given image. The problem is illustrated in Fig. 3. It can be seen from this figure that the similarity between 2 images which differ by 10 distance units, is about 82%. We, therefore, developed a custom accuracy function, which counts any prediction within 10 distance units from its true value as a true prediction. Placing the new 2D image correctly into the 3D model within such a narrow distance allowance leaves the final registration adjustment to the biologist as a fairly easy task.

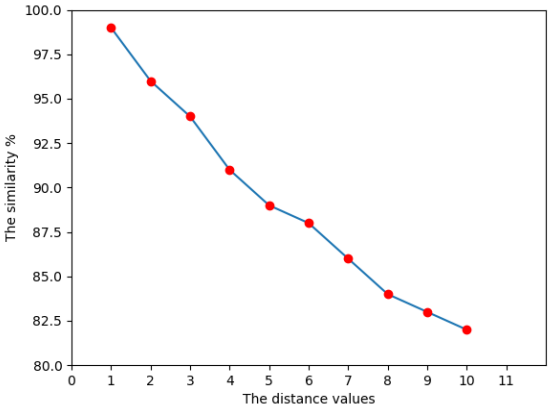

Figure 3. The change in similarity (defined by SSIM) with a change in the distance.

D. Using CNN with Feature Level Fusion

First, we analyzed the multimodal data most straightforwardly by using a CNN with two inputs, one for the image data and the other for the anatomical textual data, and merging these different types of data as shown in Fig. 4. The model consisted of another CNN called sub-CNN for the preprocessing of the image dataset, which was used to extract the dominant and relevant features that were later concatenated with the textual dataset. This model had moderate performance, the prediction accuracy was 83.20%, and the MAE was 8.7133.

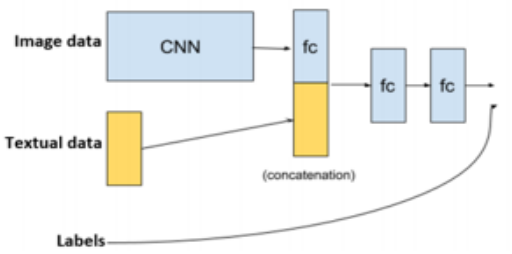

Figure 4. Illustration of CNN with a fusion of different features together.

E. Stack Generalization Ensemble for Locating the Image

To improve the accuracy of prediction, we used a stack generalization algorithm to deploy different machine learning methods for different data modalities.

1) The overall structure of the algorithm

The developed stack generalization algorithm consists of two conceptual parts, called Base models and a Meta learner model. The main aim of the Base models is an analysis of one specific type of data, i.e., either image data or textual data. Then, outputs of different Base models are combined by the Meta-learner model to generate the final prediction. We tested different Base models to optimize analysis for each data modality separately and then conducted experiments to find the best Meta-learner model. Fig. 5 presents the schematic of the overall architecture of the algorithm. The accuracy function used to calculate prediction accuracy is described in the experimental setup section.

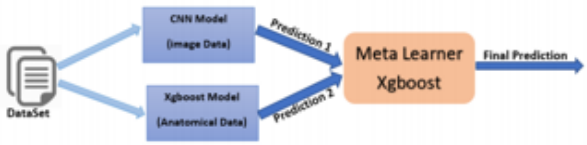

Figure 5. The overall scheme of the algorithm.

2) The base model for the image data

To build a predictive image-based model, we used different popular deep learning architectures, such as VGG16 [19], VGG19 [19], InceptionV3 [20], [21], and DenseNet, as well as a smaller custom CNN model. We initialized VGG16, VGG19, InceptionV3, and DenseNet with weights trained using the ImageNet dataset [22]. To make these models work for our data, we froze the first few layers of these models and re-trained the rest of the layers. We assumed that the first layers of these models have weights trained on large datasets to extract basic features efficiently. Thus, re-training only the last layers adjusted analysis of basic features to our specific dataset and speeded up the training process. Table I compares the performance of the CNN models during training and validation. Our custom CNN model and VGG16 had the lowest validation loss, while InceptionV3 and DenseNet had the highest validation loss. These results indicate that smaller models perform much better for our dataset than bigger and much more complicated models.

©2022 Journal of Image and Graphics

TABLE I. COMPARISON OF THE PERFORMANCE OF DIFFERENT CNN IMAGE-BASED MODELS

Models	Training loss	Validation loss
VGG16	11.5	11
VGG19	16	15.3
InceptionV3	22	21.7
DenseNet	22.5	21.5
Custom CNN	7.9	7.6

We also calculated the accuracy for the best-performing CNN architecture and obtained an accuracy of 97.2% based on the accuracy function described before. Fig. 6 shows some of the CNN model prediction results.

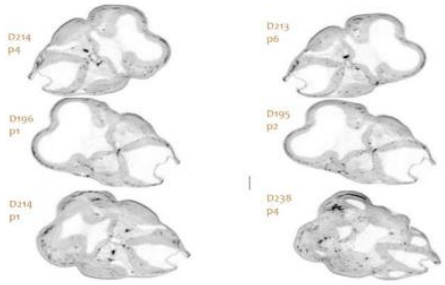

Figure 6. The pictures on the left are the query images for the CNN image model results, and the right is the matched images.

3) Base model for textual data

For analysis of textual data, we tested the CNN and the two-following tree-based machine learning methods: XGBoost [12] and Random Forest (RF) [13]. The performance of the CNN model, in this case, was unsatisfactory, and we mainly focused on the tree-based methods. The XGBoost model was trained as a regressor with 10-fold cross-validation and two stopping criteria to prevent over-fitting, such as maximum tree depth and a maximum number of gradient boosted trees. We conducted a grid search for various parameter configurations to achieve the optimum combination of the method hyperparameters. The best performing hyperparameter set for the XGBoost model is presented in Table II. The performance for this set of values gave an MSE = 417.

TABLE II. OPTIMAL HYPERPARAMETERS FOR THE XGBOOST MODEL

Parameter	value
Maximum tree depth for base learners (max_depth)	8
Number of gradients boosted trees (n_estimators)	1000
The minimum sum of instance weight needed in a child (min_child_weight)	1
Subsample ratio of the training instance (subsample)	1
Subsample ratio of columns when constructing each tree (colsample_bytree)	0.4

Also, to find optimal hyperparameters for the Random-forest regressor, we used grid search. Table III represents the optimal set of hyperparameters that provided an MSE=436. Table IV shows the accuracy of the XGBoost and the Random Forest models.

TABLE III. OPTIMAL HYPERPARAMETERS FOR THE XGBOOST MODEL

Parameter	value
Number of trees	400
Max tree depth	15
Number of features	Sklearn (python) 'auto.'

TABLE IV. COMPARISON OF THE ACCURACY OF XGBOOST AND RANDOM FOREST METHODS FOR THE ANALYSIS OF THE TEXTUAL DATA

Models	**Testing accuracy**
Random Forest	49.5%
XGboost	47%

4) Meta learner model

The main aim of the Meta learner model is to combine predictions from the two Base learner models and generate a final prediction (Fig. 3). To find the best model for the Meta learner, we tested Random Forest, XGBoost, and CNN models trying to find the optimal set of hyperparameters for each model. The XGBoost model with the same set of hyperparameters as the Base model analyzing textual data (see Table II) gave the best performance. Fig. 7 and Fig. 8 represent the dependence of training and testing accuracy on the number of iterations for XGBoost and Random Forest meta learner models. Table V shows the comparison of model performances for different learner models. It can be clearly seen from Table V that using stack generalization ensemble for analysis of multimodal data with tree-based Meta learner models outperforms both CNN with a fusion of different features and the individual models for each data modality.

TABLE V. COMPARISON OF PERFORMANCE OF DIFFERENT MODELS FOR THE PROBLEM

Meta Learner Models	Training MAE	Testing MAE	Testing Accuracy
Random Forest (meta learner)	2.19	2.24	99.37%
XGboost (meta learner)	1.60	1.84	99.55%
CNN Image only	6.41	6.38	97.2%
CNN feature Fusion	8.7133	8.7133	83.20%

Figure 7. The training and testing accuracy depend on the XGBoost base model for textual information on the number of iterations.

Figure 8. The dependence of training and testing accuracy for Random Forest-based model for textual information on the number of iterations.

Here are some examples (Fig. 9-Fig. 14) of the queries and their location prediction made by the Multimodal designed for predicting the image location([distance, pitch]):

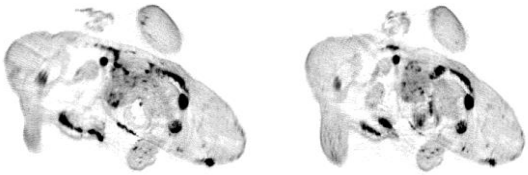

Figure 9. Image [392, 7] predicted as [395, 10].

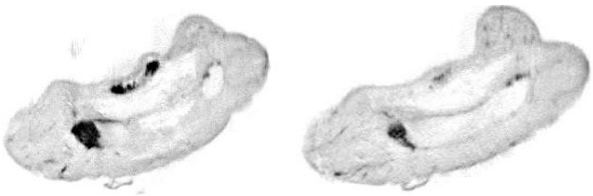

Figure 10. Image [382, 72] predicted as [390, 74].

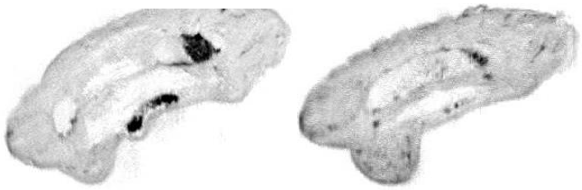

Figure 11. Image [378, 73] predicted as [390, 76].

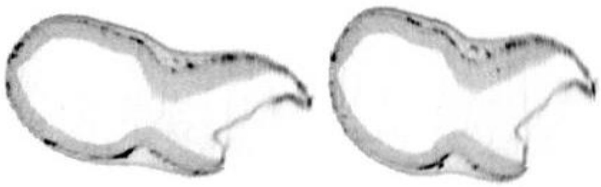

Figure 12. Image [132, 50] predicted as [135, 49].

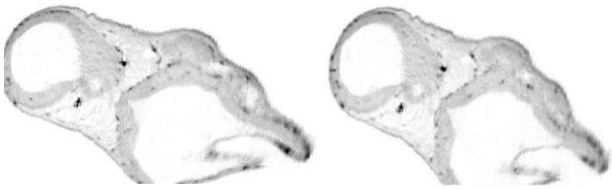

Figure 13. Image [204, 44] predicted as [203, 43].

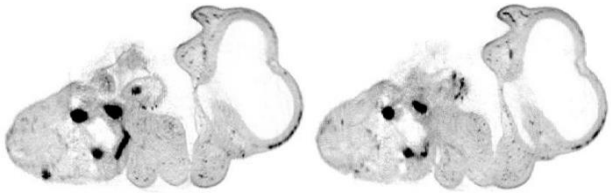

Figure 14. Image [308, 48] predicted as [304, 48].

III. Result and Discussion

The recent developments in deep learning, specifically Convolutional Neural Networks (CNN), have considerably increased the performance of machine learning methods in various computer vision tasks, Such as the analysis of medical imaging information. However, the task of 2D to 3D image registration is still a challenge for CNN methods. In our case, the complexity of the task is even higher because of the structural similarity between consecutive 2D slices of a 3D model in EMOUSE Atlas and the limited amount of data. Thus, it is not surprising at all that the straightforward application of CNNs to the multimodal data resulted in moderate accuracy. We proposed a novel image and text-based multimodal approach to address the problem, which utilizes a stack generalization approach to combine the best features of different machine learning methods into a single pipeline. The main idea of this method is to find the best algorithms for each data modality separately and then combine the predictions using a novel Meta learner model. We analyzed 2D image slices from the EMAP dataset using CNN models and textual anatomical information using tree-based models. We used 5 variations of CNN models for image analysis and 2 different tree-based models for textual analysis to find the best base model for each particular task. For analysis of 2D image slices obtained from the EMAP dataset, we tested 5 different CNN models. The loss during training and validation phases of the CNN models is reported in Table I. We found that more complex CNN models performed worse than simpler ones. The lowest training and test losses were observed for our custom CNN model and for the VGG16 model. We suggest that the bad performance of complex CNN models in our case is explained by the small size of the dataset used.

For the analysis of textual data, we used 2 different tree-based models, Random Forest and XGBoost. We also tested a CNN model for this data modality. However, although the CNN model performed well for image data, the accuracy of the CNN for the analysis of textual data was significantly lower than for tree-based methods. We found optimal hyperparameters for each tree-based model using grid search. Comparison of model performances showed that the Random Forest performed slightly better than XGBoost. We tested different models to combine the results of the image and textual model together and generate the final prediction.

IV. Conclusion

Our tests showed that the resulting stack generalization algorithm outperforms both CNN, which combines all data

©2022 Journal of Image and Graphics

modalities and individual models for each data modality when tree-based models are used as a Meta learner model. It is worth noting that although the accuracy of the image Base model (97%) was much higher than the accuracy of the textual Base model (49.5%). In summary, we have shown that combining CNNs with tree-based methods using stack generalization results in a powerful prediction tool for the image registration tasks when the size of the dataset is limited.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Bassam Almogadwy is the primary researcher of the work and main author of the paper. Albert Burger and Nick Taylor have contributed expertise in biomedical atlasing and Machine Learning, respectively. They also acted in a supervisory role. Albert Burger and Nick Taylor have made contributions as authors, primarily during revisions of the paper. All authors have approved the final version.

References

[1] N. Chuang, *et al.*, "An MRI-based atlas and database of the developing mouse brain," *NeuroImage*, vol. 54, pp. 80-89, 2011.

[2] M. Fernández-Delgado, M. S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, and M. Febrero-Bande, "An extensive experimental survey of regression methods," *Neural Networks*, vol. 111, pp. 11-34, 2019.

[3] S. K. Motsinger, "Complete anatomy," *Journal of the Medical Library Association*, vol. 108, pp. 155-157, 2020.

[4] R. B. Puchalski, *et al.*, "An anatomic transcriptional atlas of human glioblastoma," *Science*, vol. 360, p. 660, 2018.

[5] L. Richardson, *et al.*, "EMAGE mouse embryo spatial gene expression database: 2014 update," *Nucleic Acids Research*, vol. 42, pp. D835-D844, 2013.

[6] A. A. Goshtasby, *Image Registration: Principles, Tools and Methods*, Springer Publishing Company, Incorporated, 2012.

[7] B. Zitová and J. Flusser, "Image registration methods: A survey," *Image and Vision Computing*, vol. 21, pp. 977-1000, 2003.

[8] E. Ask, O. Enqvist, L. Svärm, F. Kahl, and G. Lippolis, "Tractable and reliable registration of 2D point sets," presented at European Conference on Computer Vision, 2014.

[9] X. Cao, J. Fan, P. Dong, S. Ahmad, P. T. Yap, and D. Shen, "Image registration using machine and deep learning," in *Handbook of Medical Image Computing and Computer Assisted Intervention*, S. K. Zhou, D. Rueckert, and G. Fichtinger, eds., Academic Press, 2020, pp. 319-342.

[10] J. V. Hajnal and D. L. Hill, *Medical Image Registration*, CRC Press, 2001.

[11] A. Li, S. Luo, Y. Liu, and H. Yu, "The equivalency between a decision tree for classification and a feedback neural network," presented at 7th International Conference on Signal Processing, 2004.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," arXiv e-prints, arXiv:1603.02754, 2016.

[13] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[14] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy and Buildings*, vol. 147, pp. 77-89, 2017.

[15] M. Fernández-Delgado, M. S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, and M. Febrero-Bande, "An extensive experimental survey of regression methods," *Neural Networks*, vol. 111, pp. 11-34, 2019.

[16] T. F. Hayamizu, M. N. Wicks, D. R. Davidson, A. Burger, M. Ringwald, and R. A. Baldock, "EMAP/EMAPA ontology of mouse developmental anatomy: 2013 update," *Journal of Biomedical Semantics*, vol. 4, p. 15, 2013.

[17] M. D. Bloice, C. Stocker, and A. Holzinger, "Augmentor: An image augmentation library for machine learning," arXiv preprint arXiv:1708.04680, 2017.

[18] Y. Shima, "Image augmentation for object image classification based on combination of pre-trained CNN and SVM," *Journal of Physics: Conference Series*, 2018.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv e-prints, arXiv:1409.1556, 2014.

[20] C. Szegedy, *et al.*, "Going deeper with convolutions," arXiv e-prints, arXiv:1409.4842, 2014.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," arXiv e-prints, arXiv:1512.00567, 2015.

[22] O. Russakovsky, *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211-252, 2015.

Copyright © 2022 by the authors. This is an open access article distributed under the Creative Commons Attribution License (CC BY-NC-ND 4.0), which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.

Bassam Almogadwy has a master's degree in Artificial Intelligence from the University of Manchester. He is currently a PhD student at Heriot-Watt University. His current research focuses on using deep learning techniques to map 2D images to 3D volume in Biomedical Atlases.