Ensemble of Deep Transfer Learning Models for Parkinson's Disease Classification

Kiranbabu Rajanbabu, Iswarya Kannoth Veetil, V. Sowmya, E. A. Gopalakrishnan, and K. P. Soman

Abstract Diagnosis is the key step forward to cure a disease. Deep learning is becoming popular as a tool for usage in medical diagnosis. The existing literature using deep learning for the diagnosis of Parkinson's disease (PD) by transfer learning of MRI data was limited to the AlexNet architecture. The present work aims to inculcate commonly used deep learning architectures using transfer learning for effective diagnosis of PD using MRI data. The best three performing models are selected based on the standard metric called F1-score. An ensemble model is proposed based on the maximum probability across all the selected models for PD classification. The approach mainly focuses on the effective diagnosis of PD. The performance of the proposed ensemble approach is validated using the standard metrics known as F1-score and classification accuracy. Among the commonly used deep learning architectures, it was found that VGG19 is better than the existing state-of-the-art, which is AlexNet. The proposed ensemble approach applied on top three commonly used deep learning models led to the improvement in accuracy to 0.978. The false positive (FP) and false negative (FN) were reduced significantly by the proposed ensemble approach.

Keywords Deep learning · Transfer learning · Ensemble method · Parkinson's disease · CNN

1 Introduction

Parkinson's disease (PD) is a neurodegenerative disorder that affects predominantly dopamine-producing neurons in a specific area of the brain called substantia nigra [1]. People with PD may experience tremors—mainly at rest and described as pill rolling tremor in hands and other forms of tremors, limb rigidity, gait and balance problems [2]. However, these symptoms may vary from person to person. These

K. Rajanbabu (✉) · I. K. Veetil · V. Sowmya · E. A. Gopalakrishnan · K. P. Soman
Centre for Computational Engineering and Networking, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India

V. S. Reddy et al. (eds.), *Soft Computing and Signal Processing*, Advances in Intelligent Systems and Computing 1340, https://doi.org/10.1007/978-981-16-1249-7_14

damaged neurons or structural changes in the brain can be mapped using magnetic resonance imaging (MRI) [3].

In the literature, artificial intelligence (AI) helps in pattern recognition through features extracted from the input MRI data [4]. Considering the improvements in AI and machine learning (ML), it is much easier to infer features, and it is appropriate to make use of AI and ML to help human life. In the literature, the work done with minimal preprocessing on PD classification using MRI data is limited. Sivaranjini et al. [5] proposed the deep transfer learning of AlexNet model on T2-weighted MR images. An accuracy of 88.9% is reported using the transfer learning approach. As the commonly used deep learning architectures other than AlexNet are not explored for the PD classification using MRI images, the present work focuses on analyzing the performance of the commonly used deep learning architectures.

So, in this work, we have proposed the transfer learning approach on commonly used deep learning architectures to detect Parkinson's disease. Further, an ensemble model is proposed to determine the class labels. The class label is determined based on the maximum predicted probability computed from top three commonly used deep learning architectures. The top three deep learning models were selected based on F1-score. This approach avoids the problem of hard classification. It also takes the advantage of features extracted from multiple deep learning architectures used commonly for image classification.

2 Methodology

2.1 Architectural Performance

All the architectures used in this methodology are obtained from Keras. The model used in this study is pre-trained on a large image dataset called ImageNet [6], which contains 1000 categories. None of the layers are frozen during training. Also, the features learned in the ImageNet [6] model is customized to Parkinson's disease by retraining the architecture. The main motivation behind using the pre-trained models is the feasibility for the convergence models due to the pre-trained weights in comparison with custom models, which starts with the random weights [7]. Additional dense layers are added after the final MaxPooling layer in the current work to customize the commonly used architectures, specific to Parkinson's disease use case.

All the model architectures used in this study have an input size of 224 * 224 * 3. The proposed modification in the commonly used deep learning architectures for PD classification is shown in Fig. 1. In order to remove the last layer from Keras, "*include Top* = *False*" was added when importing the pre-trained architecture from Keras. It is then connected with a GlobalMaxPooling layer to reduce the dimension of the data. Each architecture has a different shape in the last layer, and hence, including the GlobalMaxPooling (GMP) down-samples the dimension of the data

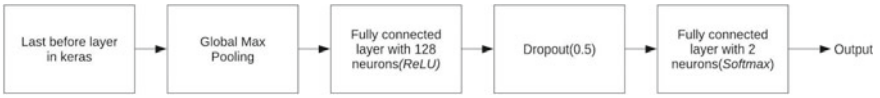

Fig. 1 Proposed modifications in the commonly used deep learning architectures for PD classification

without hardcoding the fully connected layer. The MaxPooling improves the translation invariance of the model. The MaxPooling layer is then connected to a fully connected layer with 128 neurons and rectified linear unit (*ReLU*) as an activation function. This fully connected layer helps in transition of the data from MaxPooling to Softmax. The fully connected layer holds the aggregated information from all the layers in architecture that are of prime importance.

The fully connected layer is then connected with a dropout of 0.5. Large neural nets can often overfit the data which leads to poor performance of the model, and the model will not be generalizable. Dropout is a regularization technique that approximates training a large neural network by randomly dropping few neurons [8]. Dropout rate is chosen as 0.5 which translates into 50% of the fully connected layers' output being dropped out or ignored and then connected to the Softmax layer. Softmax layer is a squashing function that limits the output range from 0 to 1. This allows the output to be interpreted as the predicted class probability distribution.

In the present work, the implementation of VGG16 architecture [9] modified for Parkinson's disease classification is explained as follows: VGG16 on removing the last layer as specified contains 7 * 7 * 512 layer. This layer is then connected with a GlobalMaxPooling layer, which helps to reduce the dimension of data to 1 * 1 * 512. It is then fed as input to the fully connected layer with 128 neurons and *ReLU* activation (to avoid overfitting) [10], which produces the data of dimension 1 * 1 * 128. A dropout layer is added with dropout rate of 0.5, and a Softmax layer is added to obtain the classification output in the shape 1 * 1 * 2. Figure 2 shows the changes made with Keras VGG16 used in the present work. The same changes are followed for all the architectures used in this study.

2.2 Ensemble Method

In the proposed ensemble approach, the top three models are chosen based on weighted average F1-score. The architectural models are limited to top three because including multiple architectural models might degrade the performance of the ensemble method. The test data is passed through each model, and the final prediction score for each class is obtained. The prediction scores are computed as shown in Fig. 3 (proposed approach). The class label of the test data is determined based on the maximum probability value obtained from the top three models. The proposed ensemble approach leads to increase in precision and decrease in the misclassification error.

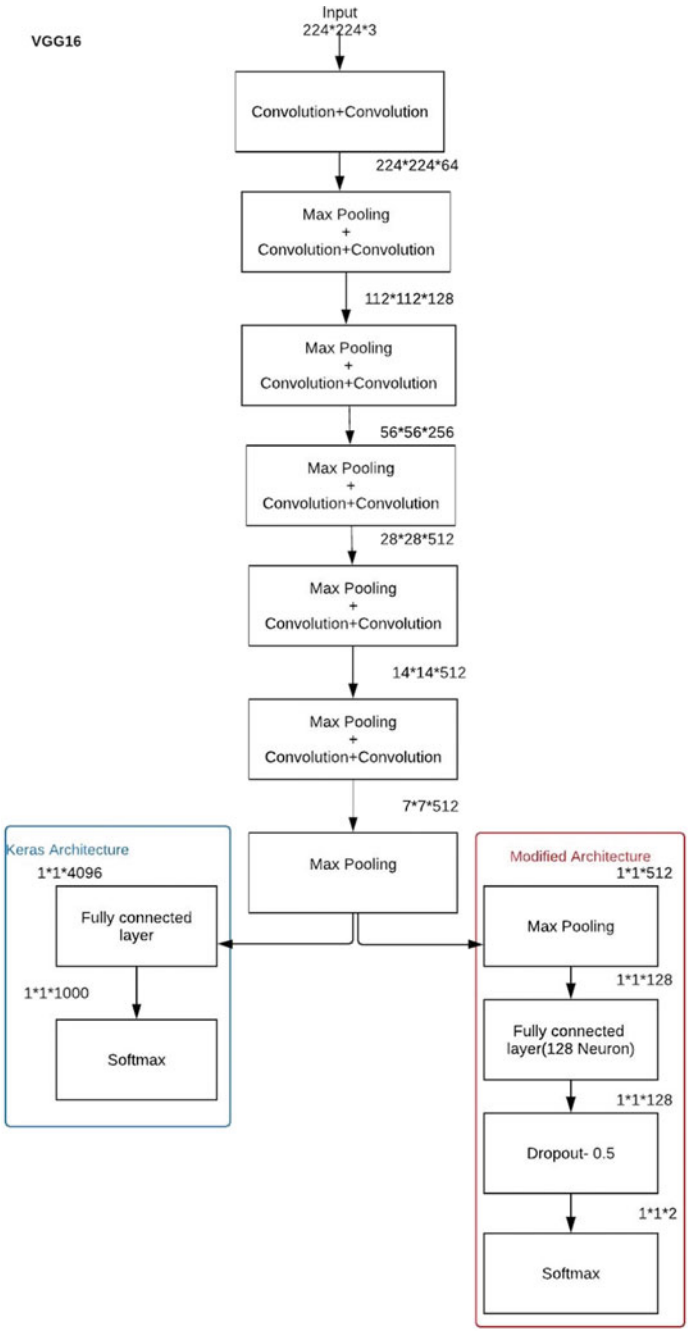

Fig. 2 Proposed modified architecture of VGG16 used for PD classification

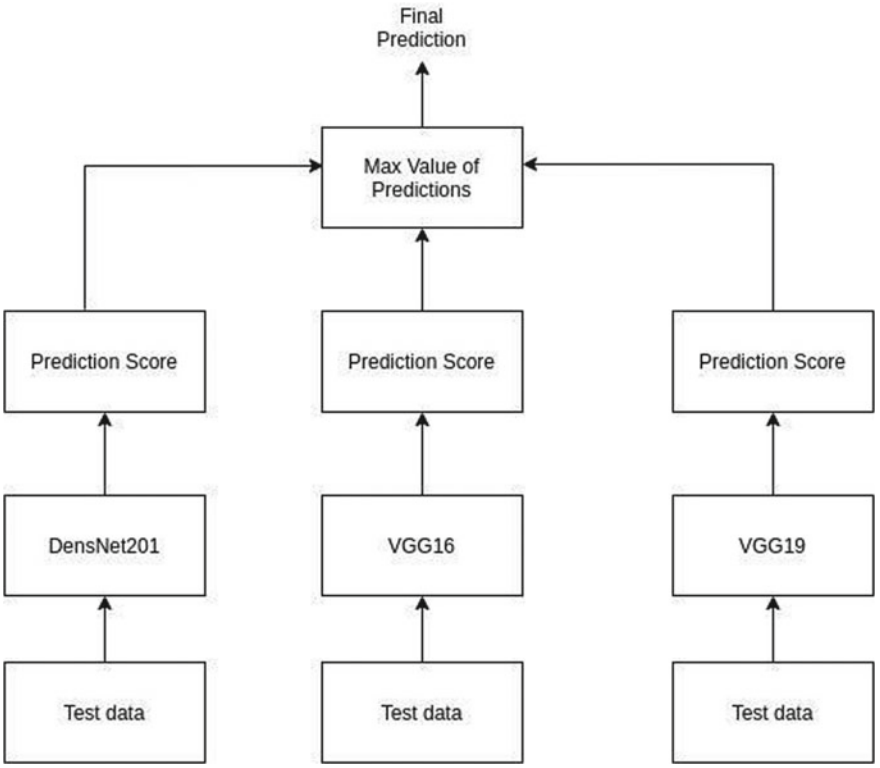

Fig. 3 Proposed ensemble approach using the top three commonly used deep learning architectures for PD classification using MRI images

2.3 Dataset

Data used in this work is obtained from the Parkinson progression marker initiative (PPMI) [11]. The PPMI dataset consists of multiple categories of data. We have focused on MRI in this study.

The PPMI data used in the present work consists of T2-weighted MR images. These images are acquired using 3T Siemens Trio Tim Scanner in axial orientation, which gives us 150 PD and 92 age matched Normal Cohorts(NC). All the images were present in DICOM format with 48 slices per image. The experiment is carried out with 80% of data for training and rest 20% for testing. The number of images used for training, validation and testing are 12,800, 3200 and 4000, respectively.

The learning rate chosen by Sivaranjini et al. [5] was 0.001. However, the learning rate is fixed at 0.0001 with a reduction by 10%. This is because, lower the learning rate, better would be the loss convergence of the model. Raw data obtained from the PPMI [11] dataset has an unequal number of images for each class. In order to

Table 1 Number of images used in the original and the proposed work

Name	PD	Normal
Original dataset	7205	4419
Modified dataset (proposed work)	10,000	10,000

counter the class imbalance, we performed data augmentation, which leads to an equal number of images for both normal and PD classes (tabulated in Table 1).

The following data augmentation techniques were used to tackle the class imbalance: Image flip left/right, image flip top/bottom, rotation 10-degrees (left/right) and Gaussian blurring with 0.8 mean and standard deviation (0,1). Thus, a total number of 20,000 images were obtained. Out of the available number of images, 16,000 images (80% of data) are used for training with 8000 from each class, and also, a subset of training data 3200 images (20% of training data) are used for validation. The remaining 4000 images (20% of data) are used for testing.

3 Results and Discussion

3.1 Experimental Results from Commonly Used Deep Learning Architectures

Each model is trained for 50 epochs. The number of epochs is fixed based on the saturation of the validation loss curve of the model. So, each epoch is saved periodically. It is tough to do a manual model evaluation. Therefore, the metric chosen for the model evaluation is accuracy and F1-score [7]. The reason for choosing F1-score as a metric is that it captures the false positive (FP) and false negative (FN) unlike accuracy, which captures only the true positives and true negatives. Also, in a medical scenario considering the life of human beings at stake, we need to have minimum FP and FN (Table 2).

The AlexNet model discussed by Sivaranjini et al. [5] shows an accuracy of 0.889. However, we can infer that VGG19, VGG16 and DenseNet201 perform better when compared to AlexNet. The VGG16 is able to detect more number of positive cases, when compared to the other two models. Thus, this model can detect persons with Parkinson's disease effectively but also misclassifies the normal case as affected

Table 2 Performance of the commonly used deep learning architectures for PD classification based on accuracy and F1-score

Architecture	Accuracy	F1-score normal	F1-score PD
VGG16	0.917	0.913	0.921
VGG19	0.926	0.923	0.929
DenseNet201	0.902	0.898	0.905
MobileNetV2	0.867	0.863	0.871

by PD, which is clear from the number of false positives. VGG19 has less false positives in comparison with VGG16. VGG19 can differentiate the normal case from PD effectively, i.e., it is able to detect normal cases better when compared to VGG16. DenseNet has less true negatives for PD, in comparison with both VGG 16 and VGG 19. This helps in the reduction of the misclassification of PD data. The advantage of each of the top three models motivated the proposed ensemble approach to increase the classification accuracy of PD and reduce the misclassification error.

3.2 Experimental Results from the Proposed Ensemble Model

We have chosen models for the proposed ensemble approach based on weighted F1-score. The model for ensemble approach is VGG19, VGG16 and DenseNet201. The confusion matrix obtained using the proposed ensemble approach for PD classification is given in Table 3. The performance metrics for the top three model used in the present work and the proposed approach are given in Table 5.

The probability of prediction for an actual test data along with the class label is given in Table 4. For example, the predictions on the second row as shown in Table 4 for PD are very similar to normal in the case of DenseNet. This is called hard classification. But, when the same data is passed through multiple models, there is a possibility that different models may capture the class specific features that are uncorrelated to the difference in the architectures of the models. VGG16 is able to detect the label of the class as normal with 0.99 predicted probability. Hence, by using

Table 3 Confusion matrix obtained using the proposed ensemble approach for PD classification using MRI data

1957	43
45	1955

Table 4 Architectural models with prediction score converted to CSV

Image	Dense Net Pd	Dense Net normal	VGG 19 PD	VGG19 normal	VGG 16Pd	VGG16 normal	Label
1	0.999998	1.88 e−06	1.00 e+00	4.87 e−08	9.99 e−01	0.000004	PD
2	0.581974	0.41029	9.02 e−01	9.78 e−02	0.999 759	0.000241	PD
3	0.999951	4.85 e−05	1.00 e+00	3.16 e−08	9.99 e−01	0.000220	PD
4	0.225530	7.74 e−01	1.25 e−08	1.00 e+00	4.52 e−04	0.999548	Normal
5	0.000065	9.99 e−01	7.85 e−06	9.99 e−01	7.54 e−06	0.999992	Normal

Table 5 Performance metrics obtained using the top three models and the proposed ensemble approach

Model	Accuracy	Precision	Recall
VGG16	0.917	0.957	0.935
VGG19	0.926	0.957	0.889
DenseNet201	0.902	0.932	0.867
Ensemble method	0.978	0.979	0.978

the proposed ensemble method, we are also able to remove the hard classification problems. The class label corresponding to the maximum prediction score from the top three models leads to higher accuracy, less false positives, and false negatives.

It is evident from Table 5 that false positive (FP) and false negative (FN) are reduced by greater extent in comparison with architectural models discussed in Sect. 2.1. FP contributes only 1.125% and FN 1.075% for the test data. Such low FP and FN assure that the predictions are closer to ground truth. The proposed ensemble method explained in Sect. 2.2 is able to achieve an accuracy of 97.8% which is around a 10% increase in accuracy from the benchmark result for PD classification using MRI data (accuracy: 88.9%) [5].

4 Conclusion

In the present work, the performance of commonly used deep learning architectures (VGG16, VGG19, DenseNet and MobileNet) is analyzed for PD classification using the metrics called accuracy and F1-score. The key findings of the paper are as follows:

1. The AlexNet model detects PD with 88.9 accuracy. The accuracy and F1-scores from our study show that VGG16, VGG19 and DenseNet models were able to detect Parkinson's disease more effectively than AlexNet.
2. MobileNet has lesser accuracy than AlexNet. Therefore, MobileNet is not used in the proposed ensemble approach.
3. Ensemble of VGG19, VGG16 and DenseNet201 predictions based on the maximum probability is proposed which resulted in 97.8% accuracy for PD classification, which is around 10% increase in accuracy when compared with the benchmark accuracy (88.9%).
4. The proposed ensemble approach also resulted in less FP and FN for PD class when compared to the individual performance of the top three architectures.

The future scope of the present work is to extend the proposed approach to the current pandemic COVID-19 data for effective diagnosis. The proposed ensemble approach can be extended to the architectures retrained with the COVID X-ray and normal images.

References

1. C.A. Davie, A review of Parkinson's disease. Br. Med. Bull. **86**(1), 109–127 (2008)
2. J. Obeso, M. Stamelou, C. Goetz et al., Past, present, and future of Parkinson's disease: a special essay on the 200th anniversary of the shaking palsy. Mov. Disord. **32**(9), 1264–1310 (2017)
3. S. Duchesne, Y. Rolland, M. Varin, Automated computer differential classification in Parkinsonian syndromes via pattern analysis on MRI. Radiology **16**(1), 61–70 (2009)
4. A. Marquand, M. Filippone, J. Ashburner, M. Girolami, J. Mourao-Miranda, G.J. Barker, S. Williams, P. Leigh, C. Blain, Automated, high accuracy classification of Parkinsonian disorders: a pattern recognition approach. PLoS ONE **8**(7), e69237 (2013)
5. S. Sivaranjani, C. Sujatha, Deep learning based diagnosis of parkinson's disease using convolutional neural network. Multimedia Tools Appl. 1–13 (2019)
6. O. Russakovsky, J. Deng, H. Su et al., Imagenet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015)
7. K. Radhika, K. Devika, T. Aswathi, et al., Performance analysis of nasnet on unconstrained ear recognition, in Nature Inspired Computing for Data Science (Springer, Berlin, 2020), pp. 57–82
8. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(56), 1929−1958 (2014)
9. K. Simonyan, A. Zisserman, Very deep convolutional networks for large scale image recognition (2014). arXiv preprint arXiv:14091556
10. P. Gopika, V. Sowmya, E. Gopalakrishnan, et al., Transferable approach for cardiac disease classification using deep learning. Deep Learn. Tech. Biomed. Health Inform. **285** (2020)
11. K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, et al., The Parkinson progression marker initiative (ppmi). Progr. Neurobiol. **95**(4), 629–635. https://www.ppmi-info.org/