SM2N2: A Stacked Architecture for Multimodal Data and its Application to Myocardial Infarction Detection*

Rishabh Sharma[1,2], Christoph F. Eick[2], and Nikolaos V. Tsekos[1]

MRI Lab[1] and DAIS Lab[2], Dept. of Computer Science, University of Houston, Houston TX 77004, USA
rsharma26@uh.edu, {CEick,nvtsekos}@central.uh.edu
https://www.uh.edu/nsm/computer-science/

Abstract. This work introduces a novel Stacked Multimodal(SM2N2) architecture and assess its performance in classifying whether a patient have or not Myocardial Infarction. Central to this SM2N2 architecture is the use of images and clinical data as input. Comparison studies of Multimodal Neural Network(M2N2) component of SM2N2 with AlexNet3D model demonstrated that for the small size of dataset M2N2 is faster, has less trainable parameters and results higher accuracy in this binary classification. In addition to M2N2 we also identify clinical features that are sufficient to classify normal vs pathological cases. We also train statistical models on identified clinical features and use stacking to combine outputs from statistical models and M2N2. Stacking generalizes the results and the new model learns how to best combine the results of the individual base models. One of the potential application of the M2N2 is that because of less parameters the network can be deployed on mobile devices for inference.

Keywords: MRI · heart · myocardial infarction · normal case · delayed-enhancement · classification.

1 Introduction

According to an article by CDC, every year about 647,000 deaths are caused by heart attack in the USA only [2]. Secondary to compromised coronary arteries blood flow, myocardial infarction (MI) may develop and progress into the oxygen starving myocardium. Timely diagnosis of myocardial infarction is required to identify the area affected, perform an intervention, and remove the blockage from the artery. It is often observed that years of domain expertise is required to classify patients with myocardial infarction from regular patients. Hence, making it relatively important to develop innovative methods that can

* This work was supported by the National Science Foundation award CNS-1646566. All opinions, findings, conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of our sponsors.

quickly and accurately identify the patients who are suffering from myocardial infarction. Traditionally, physicians have used DE-MRI images and clinical information together to identify the cases. However, there is very limited research in machine learning and intelligent systems that can combine multiple inputs to make predictions. In this work, we propose a novel method to evaluate whether we can identify myocardial infarction cases from normal cases by combining DE-MRI images and clinical information automatically. We also propose a stack block in this paper where we combine outputs of multiple independent models to make final decision on the data set. Our results and analysis show that our technique(M2N2) of combining multiple inputs to make classification is better than AlexNet3D that only takes single image input. Additionally, we also modified inputs for AlexNet3D to take multiple inputs and observed that M2N2 is still giving better accuracy than AlexNet3D with modified inputs on the limited low number of samples that were provided in the dataset. The challenge dataset [1] consist of 100 patients with clinical observation and DE-MRI images provided for training and testing the model.

2 Methods

2.1 System Overview

Figure 1 shows the graphical overview of our system. Future subsections will discuss the individual components in detail.

2.2 Data Preprocessing

There are twelve clinical features and we need to identify the features that are important for classification. We normalize the continuous variables(4 clinical attributes) using Z-Score normalization to change the values of columns to bring these features at a common scale, such that all the variables fall in the same range. A Pair-Plot is created on all the features replacing the continuous features with Z-Score Normalized features to identify the attributes that are able to divide labels linearly. We also fit a linear model(Ridge Regression) on all the features replacing the continuous features with Z-Score Normalized continuous features. Ridge regression shrinks slope asymptotically close to zero but will never become absolute zero. Beta/Coefficient for features that are less important start to shrink when the penalty factor is increased and after a certain number of iteration the variables that do not contribute to the model shrinks very close to zero. We remove all the variables for which we have betas/coefficients that are in the range of 0.1 to -0.1. After filtering we were finally left with six clinical features (Sex(β = 0.315), Overweight(β = 0.105), CorArtDiseaseHist(β = -0.482), ECG(β = -0.253), ZScoreNormalized_Troponin(β = -0.123), and ZscoreNormalized_Age(β = -0.124)) to use in our network and with statistical models.

Minimization term used in the ridge regression is shown in the equation 1. β represent the slope for the line or coefficients of variables and λ represents the penalty factor.

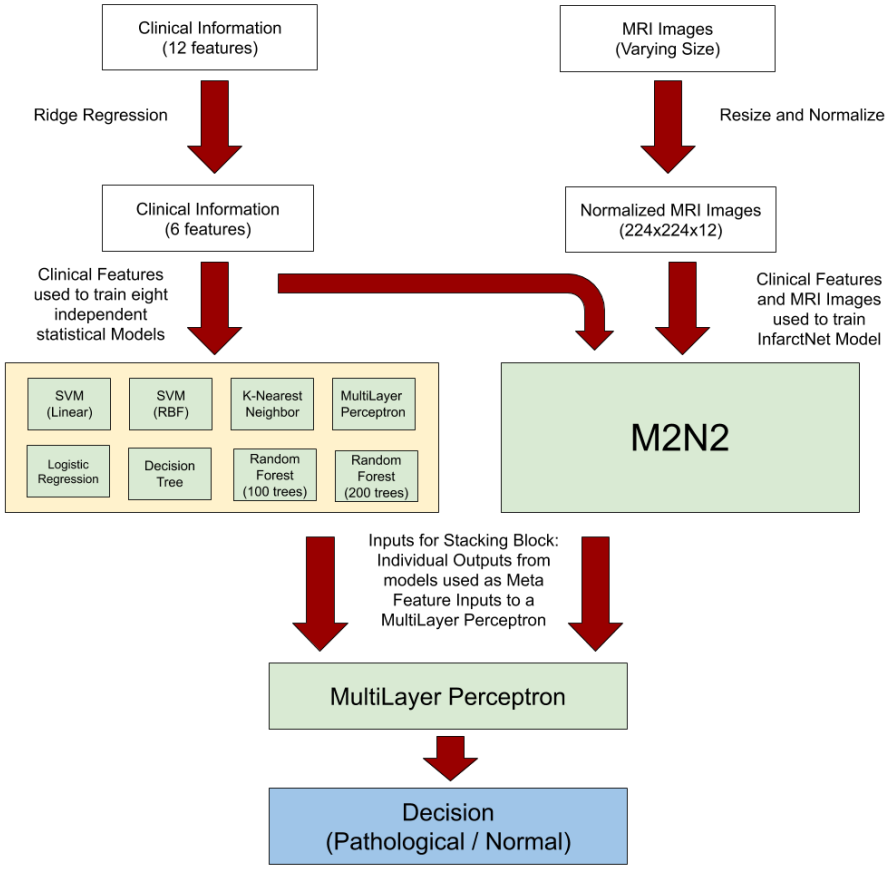

Fig. 1. SM2N2 System Overview.

$$\sum (y - y^i)^2 + \lambda \times \sum \beta^2 \tag{1}$$

Neural Networks are designed to take inputs of a constant shape, however, DE-MRI images in the challenge have inconsistent shape. DE-MRI images have a minimum of 4 slices with few having a maximum of 10 slices. To overcome the challenge of shape mismatch, we reshape our images to 224×224 and increase the total number of channels to 12 by adding zero padding. All the images are reshaped to 224×224×12 and are divided with 4096 to scale the pixels in the range of 0 to 1.

Finally, we split our data into 80% training data and 20% testing data. Training data is further split into 90% to train and 10% to validate the models.

2.3 M2N2 Architecture

We propose a novel neural network architecture described in Figure 2, that combines DE-MRI images with the clinical information. The network is a multi input, multi model neural architecture which uses Depthwise Separable Convolutional layers [3] to extract three dimensional features from images and combines it with a multi layer perceptron.

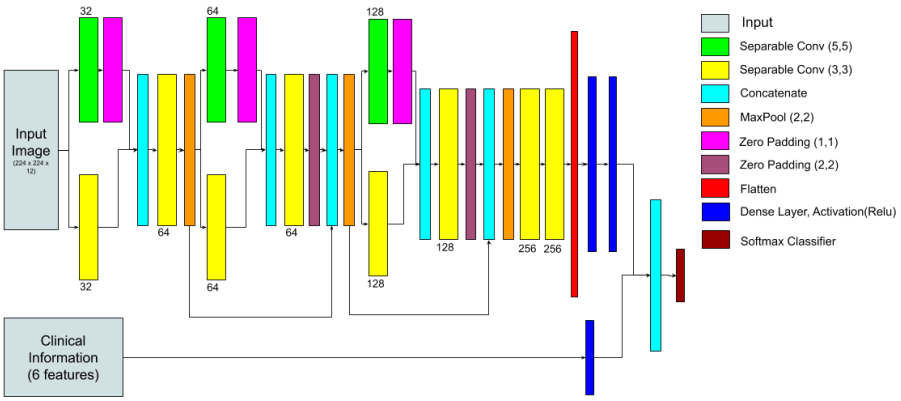

Fig. 2. M2N2 Architecture: Numbers below and above the layers represent the number of filters in each separable convolutional layer.

The first input consists of 3D images with a shape of 224×224×12. Since the images are reshaped, few of the local features are stretched, and few are shrunk down due to the original shape mismatch. Reshaping makes it difficult to use constant kernel size across the network to identify and extract infarction features. To overcome the above challenge and to capture infarction features, our architecture has separable convolutional layers with two kernel sizes. The first kernel is a 3×3 matrix, and the second kernel in parallel is a 5×5 matrix, capturing features of different sizes. The two kernels are followed with a ReLU [4] activation and batch normalization [5]. However, it raises the complication of vanishing gradients [6]. To overcome this, we use residual connections where the outputs of max pool layers and separable convolutions are concatenated to use the features from previous layers by skipping the intermediate layers. This enables the network to collect signals from max pool layer. He et al. [9] shows that using residuals can overcome the issue of vanishing gradients. The later part of the network consists of two dense networks with 1024 and 1024 neurons respectively that follows Relu Activation.

The second input consists of prepossessed clinical information with only six chosen features connected to a multilayer perceptron with 12 neurons in the hidden layer, followed with a Relu Activation.

Dense features from both the outputs are concatenated together, and the concatenated features are sent to a softmax classifier for final classification. Our architecture uses Adam [7] optimizer with a categorical cross entropy [8] loss function. We train our network for 2000 epochs with a batch size of 8.

We also train nine different statistical models on the six chosen clinical attributes to identify the possibility to classify patients without MRI scans. The statistical models are SVM with RBF kernel and linear Kernel, distance weighted K-Nearest Neighbor, Logistic Regression, Decision Tree Gini Impurity as a criterion to measure the split, Random Forest with 100 trees and 200 tress, and Multi-Layer Perceptron with 12 neuron in hidden layer.

2.4 Stacking the models

At this stage we concatenate the outputs from base models and put them together to create a set of meta features. These meta features are then used to train a multilayer perceptron with a sigmoid classifier and 18 neurons in the hidden layer to make a final decision for the class. Original class label is used as the ground truth. Multilayer perceptron used for stacking was trained for 2000 epochs with a batch size of 8. Figure Fig. 3 shows the base models and their input along with the stacking block.

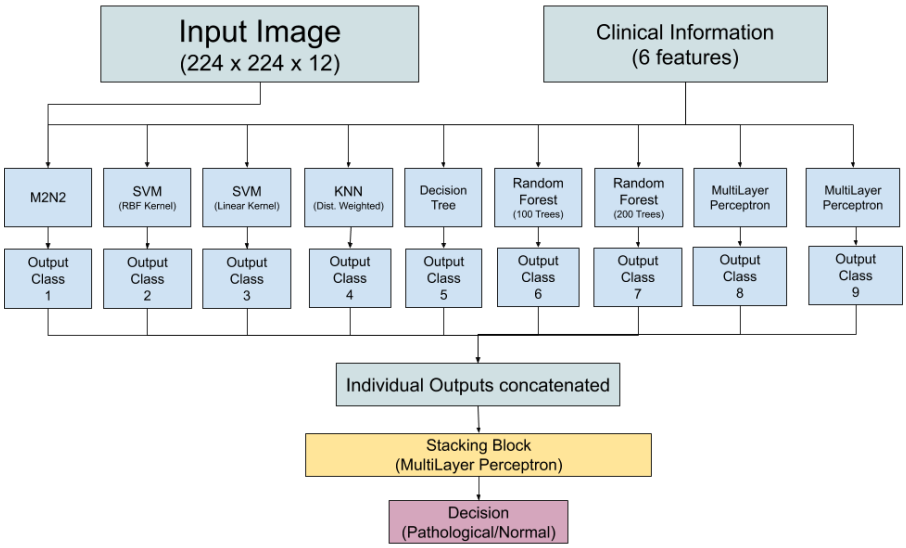

Fig. 3. Base models stacked together to choose the final class for the sample

3 Results

We compare M2N2 with AlexNet3D shown in this paper by Polat et al. [12]. We first train AlexNet3D with image inputs only and report the performance. However, the original AlexNet3D architecture is not suitable to take multiple inputs and can only take 3D images as input, because of which a comparison with M2N2 is not possible. To overcome this challenge, we modify the inputs of AlexNet3D and add a multilayer perceptron with clinical features as input. After this modification, AlexNet3D is comparable with M2N2. Figure Fig. 4 shows the modified AlexNet3D architecture with multiple inputs which is uniform with respect to the architecture of M2N2 making them comparable.

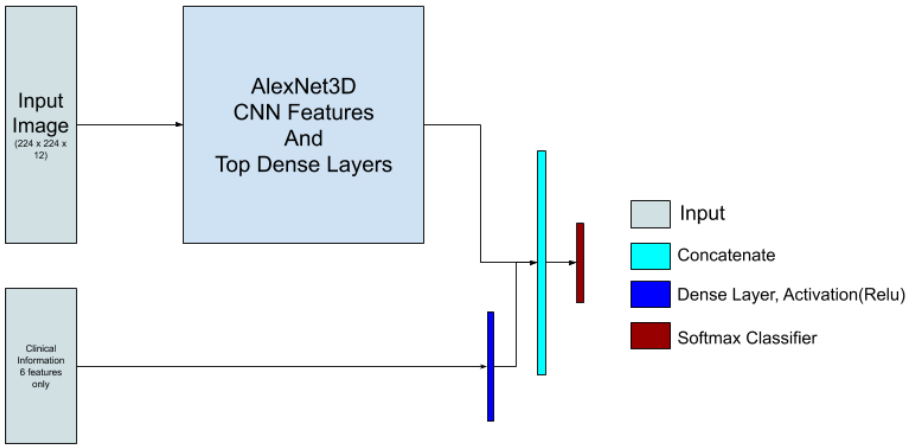

Fig. 4. AlexNet3D architecture with multiple inputs.

Table 1 represents the test and validation accuracy on the limited low number of data between AlexNet3D, AlexNet3D with multi inputs and M2N2. Table 2 compares the number of parameters for M2N2 and multi input AlexNet3D.

Table 3 represents the test accuracy after training statistical base models on clinical inputs for a limited low number of data samples. We use 10-fold cross validation on the samples and report the average test accuracy of the 10 folds. The results on statistical models show that with the limited low number of samples that are given in the dataset, it is possible to classify patients using statistical methods and clinical data alone but the M2N2 architecture has better performance then the stand alone base models. However, more research is needed on this topic which is beyond the scope of this paper.

Final classification accuracy after stacking and using base model outputs as meta feature inputs to a multilayer perceptron was 95%. Table 4 represents the confusion matrix on the limited test samples after stacking the models.

Table 1. Accuracy of models

Model	Validation Accuracy (percent)	Test Accuracy (percent)
M2N2	**87.25**	**90**
AlexNet3D (Multi Input)	83.33	80
AlexNet3D (Image Input Only)	83.33	75

Table 2. Evaluation of trainable parameters of models

Model	Total Trainable Parameters	Training time Per Epoch (seconds)
Our Model(M2N2)	**139,970,736**	**4**
AlexNet3D	318,022,232	8

Table 3. Confusion matrix score for 20 test samples on statistical models with the six clinical inputs only. Pathological cases are 'Negative' class and Normal cases are 'Positive' class

Model	Test Accuracy (percent)
Support Vector Machine (RBF kernel)	91
Support Vector Machine (Linear kernel)	90
Distance Weighted K-Nearest Neighbors	90
Logistic Regression	93
Decision Tree Classifier	87
Random Forest (100 trees)	91
Random Forest (200 trees)	92
Multi-layer Perceptron (12 hidden neurons)	87

Table 4. Confusion Matrix for 20 test samples after applying the Stacking: Positive is Normal Patient and Negative are Pathological Patients

Model	True Positive	True Negative	False Positive	False Negative
SM2N2	6	13	0	1

4 Conclusion

We create a novel stack multimodal architecture called SM2N2, which combines 3D DE-MRI images and clinical information and allows multiple inputs to a neural network. On the limited low number of samples used for training we observed that SM2N2 has an accuracy of 95%. M2N2 component of SM2N2 serves as the center for combining images and clinical information together.On the limited low number of samples used for training we observed that M2N2 has better performance compared to multi input AlexNet3D by 10%, while reducing the trainable parameters by more than 50%. Reduction in parameters has improved the training and inference time for M2N2 and, made it possible to deploy this model on mobile devices. M2N2 is inspired by Resnet [9], Inception [10] and MobileNet [11]. We also identified that it is possible to classify pathological patients from normal patients by using clinical information alone, however we can not make a conclusive statement about this finding with such a small dataset. Finally we use stacking on the meta features to generalize base models. We train

a multilayer perceptron on concatenated outputs of base models to make the final decisions. We observe that SM2N2 gives the highest accuracy. We will use SM2N2 on the final Emidec Challenge dataset.

Feature selection technique at this stage is based on statistical analysis and needs to be verified with a physician to show the clinical impact of the chosen attributes. This finding is beyond the scope of this paper and will be explored in future work.

The current performance of our network is based on just 100 patients. The dataset is extremely small and it is a challenge to analyze the performance of a network with this small dataset. Hence, we believe that more research and a large dataset is required to analyze the performance of M2N2, to determine benefits of combining images and clinical data, and compare it with other relevant architectures for performance ranking.

5 Future Work

This is an ongoing research and future work will answer some of the questions below:

- What is the advantage of techniques that combine image and clinical data versus the techniques that only take a single input?
- What is the performance of M2N2 compared to other deep learning and statistical models?
- What is the clinical influence of attributes given in the dataset and how they impact the analysis?

References

1. Lalande, A., Chen, Z., Decourselle, T., Qayyum, A., Pommier, T., Lorgis, L., de la Rosa, E., Cochet, A., Cottin, Y., Ginhac, D., Salomon, M., Couturier, R., Meriaudeau, F.: Emidec: A Database Usable for the Automatic Evaluation of Myocardial Infarction from Delayed-Enhancement Cardiac MRI. Data **5**(4), 89 (2020).
2. CDC, https://www.cdc.gov/heartdisease/facts.htm. Last accessed 17 August, 2020
3. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1251–1258. (2017)
4. Agarap, A. F.: Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375, (2018)
5. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, (2015)
6. Hochreiter, S.: The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **6**(02), 107–116 (1998)
7. Kingma, D. P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

8. Zhang, Z., Mert S.: Generalized cross entropy loss for training deep neural networks with noisy labels. In Advances in neural information processing systems, pp. 8778–8788. (2018)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 770-778. (2016)
10. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9. (2015)
11. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. (2017)
12. Polat, H., Danaei Mehr, H.: Classification of pulmonary CT images by using hybrid 3D-deep convolutional neural network architecture. Applied Sciences, **9**(5), 940 (2019)