Towards Robust Image Manipulation Detection V2 (RIMD2): Leveraging Ensemble Neural Networks and Fine-Tuned Hyperparameterisation *

Jawwad Ahmed

School of Computing, Australian National University
Canberra, Australia
u7152578@anu.edu.au

Abstract. Image Manipulation Detection is an active research problem and a point of concern as human performance in classifying manipulated from original images has been mediocre. State-of-the-art image editing applications, now available at affordable prices, have made identifying fabricated images an uphill task. In this research, we try to solve this problem using deep learning classification. We implement Ensemble Neural Networks, a popular deep learning algorithm on the eye gaze tracking dataset. We then perform extensive hyperparameter tuning and achieve an accuracy of 77.33% using our method, significantly outperforming human accuracy of 56% on the same dataset.

Keywords: Image manipulation · ensemble neural networks · eye gaze tracking.

1 Introduction

Rapid advancements in technology and increased accessibility to resources have significantly eased the real-time manipulation of images. The introduction of technologies like Deepfakes, which can mask a person's face onto another image or video in real-time, has made detecting image manipulation significantly complicated [1]. Recent developments in generative AI such as DALL-E [1] and Google's Imagen [2] have escalated this challenge. Software like Adobe Photoshop has incorporated generative AI that can produce hyper-realistic images. While these advancements in AI are impressive, they give rise to a severe problem of image credibility. This motivates us to create more robust methods for Image manipulation detection.

In this paper, we enhance the accuracy and robustness of our Robust Image Manipulation Detection model (RIMD), building upon methodologies and insights from our previous work [3]. Our prior research, "Towards Robust Image Manipulation Detection: A Comparative Study of Neural Networks and Statistical Methods vs Human Perceptions," laid a solid foundation that we aim to strengthen by implementing ensemble neural networks in our model architecture. This approach, leveraging the collective intelligence of multiple neural networks, aims to enhance model performance and improve accuracy in detecting manipulated images. The advancement signifies our commitment to evolving our models to navigate the complex challenges fabricated by image manipulation techniques.

We adopted the Eye Gaze Tracking dataset by [4] for this task. The dataset employs eye gaze tracking data collected from two infrared cameras and an IR light emitter pod, which tracks the eye movement of participants. Studies by [6] have shown that humans tend to focus on the salient features of the images. Further research by [7] indicates that luminance in certain parts of an image attracts human attention. These studies make the Eye Gaze Tracking dataset ideal for classifying manipulated images. For our study, we use a subset of this dataset, apply the deep learning technique of Ensemble Neural Networks, and perform image classification.

Human accuracy with the Eye Gaze Dataset has been poor to moderate, achieving around 56% mean accuracy [4], which is slightly better than random guessing. These limitations are attributed to participant's pre-assumptions and use of incorrect logic. However, subjects with past knowledge of image manipulation performed better in the detection task [8]. Our research employs deep learning algorithms devoid of such biases and pre-assumptions. These algorithms predict based solely on learned data patterns, potentially overcoming the above mentioned limitations.

In our previous research, the employment of shallow neural networks resulted in an accuracy of 76%[3]. Advancing further in this study, we have implemented ensemble neural networks, achieving a promising accuracy of 77.33%. Additionally, we have continued to use decision trees and maximum likelihood techniques from our previous research as complementary approaches, enabling a comprehensive evaluation of the robustness and generalizability of our findings."

Section 2 details the methodology implemented and various techniques employed for the research. Section 3 and 4 presents the result, experimentation and discussion. Finally, section 5 concludes the paper and outlines future research pathways.

* Supported by School of Computing, The Australian National University.

2 Method

2.1 Dataset Description and Preprocessing

We utilize the Eye Gaze Dataset from [4]. The dataset was formed with the participation of 80 volunteers from diverse backgrounds, having a mean age of 24.4 and a standard deviation of 8.7. Test images were manipulated using techniques such as cloning, repositioning, and splicing (see Fig. 1). Volunteers were shown both original and manipulated images and provided with prior information on how images can be manipulated. Each image was accompanied by a set of questions, and the responses were recorded, and the authors performed a Glaserian grounded theory analysis on these responses [9].

For data collection, each participant's eye gaze was recorded using dual infrared cameras and an IR light emitter pod. The dataset records the fixation of eye gaze on the regions of images [4].

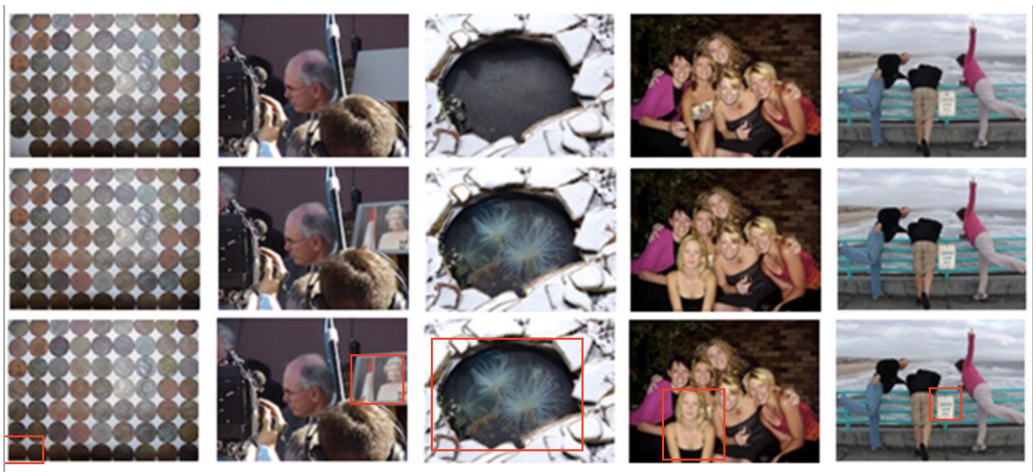

Fig. 1. Comparison of Original and Manipulated Images. Row 1 shows original images, row 2 illustrates manipulated images and the final row shows the manipulated part of the image in red bounding box

For our research, we selected a subset of the Eye Gaze Dataset. A detailed analysis revealed that the dataset is well-balanced, containing a nearly equal distribution of manipulated and unmanipulated images, as illustrated in Fig. 2. This balance allowed for a more reliable and unbiased classification process, requiring only minor adjustments and normalization. Our subset consists of 373 data points. Each data point is characterized by feature vectors such as the number of fixations, duration of fixation, number of fixations on the manipulated part of the image, and the duration of the same. These features are crucial as they provide insights into the participants' focus areas and engagement levels with each image, aiding in the manipulation detection process. To ensure a consistent scale across features and to mitigate the influence of outliers, we normalized the data by removing the mean and scaling each feature to unit variance.

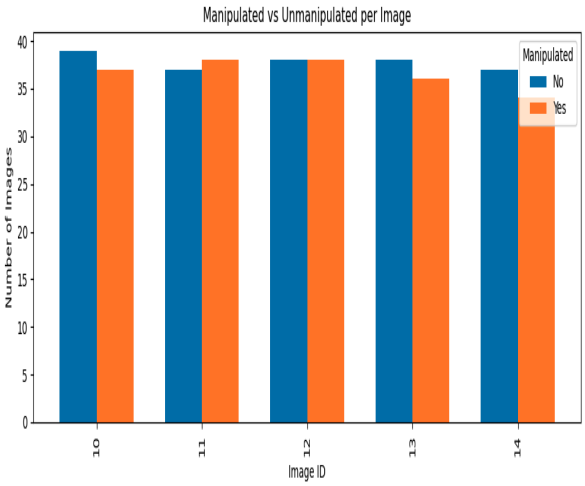

Fig. 2. Distribution of Manipulated and Unmanipulated Images

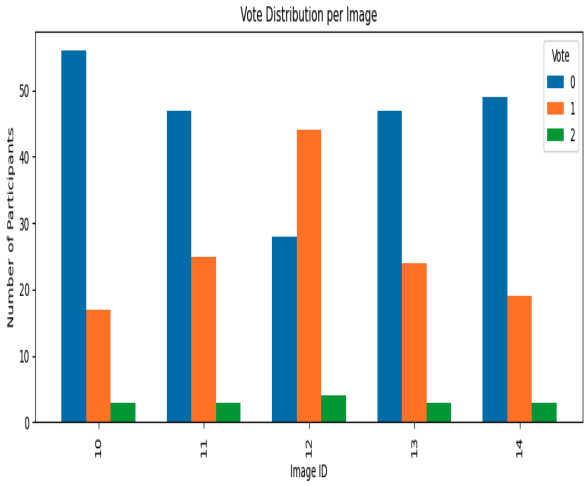

Fig. 3. Participants Voting Patterns on Image Manipulation

Upon closer examination of Fig. 3, we observe the voting patterns of participants when analyzing each of the five dataset images. Votes are categorized as: '0' for unmanipulated, '1' for manipulated, and '2' for inconclusive. A brief analysis reveals a consistent challenge for participants in identifying manipulated images, concluding that humans are bad at detecting manipulated images. A more nuanced exploration is facilitated by Fig. 1, which describes the specific manipulations in each image. The first image reveals an added coin, the second image exhibits an inserted picture of Queen Elizabeth, the third image displays additional artwork at the center in the water,the fourth image features an extra person in the front, and the fifth image has added text on the signboard. The variability in detecting manipulated images is poor across the dataset, with 'image12' emerging as an outlier due to more noticeable manipulations in the water region, facilitating easier detection. In the next section, we implement the deep learning technique of ensemble neural networks and perform extensive hyperparameter tuning, making the model more robust and accurate for image manipulation detection.

2.2 Ensemble Neural Network Classification

Ensemble neural networks are the deep learning solution to improve generalization over small datasets[5]. It is a process of combining multiple neural networks on the dataset and using their collective predictive power in making predictions, as seen in Fig 4. We have used a shallow three-layer neural network in [3], achieving an accuracy of 76%. We will refer to this as a base architecture. Our ensemble neural network takes this a step further by aggregating the power of three neural networks and making predictions. We have used the mean of the predictions from all three models. We utilize the ReLU activation function for the hidden layers and a Sigmoid function for the binary classification task.

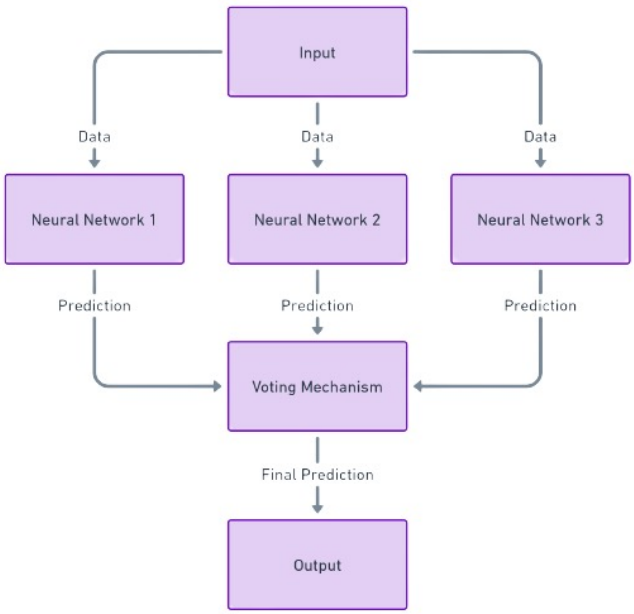

Fig. 4. Illustration of the Ensemble Neural Network architecture.

The feature columns were transformed into tensors suitable for model input. We opted for a mini-batch of size 64 for training, and for testing the base model, the entire test set was processed in a single batch. Given the binary nature of our classification task, we chose the binary cross-entropy loss for loss computation. We implemented Adam optimizer to optimize the model with a learning rate 0.001. A weight decay of 0.0001 was introduced to mitigate overfitting, given the small size of our dataset.

All neural networks are trained for a maximum of 1000 epochs. After each epoch, we evaluated the model's performance using the test set's F1 score and accuracy metrics. The model with the highest accuracy is deemed the best and subsequently used for testing. We employed a standard threshold of 0.5 to predict whether an image was manipulated or original. Finally, we generate a confusion matrix that provides the model's classification capabilities.

$$avg\_predictions = \frac{1}{N}\sum_{i=1}^{N} predictions_i \tag{1}$$

We have used the mean of predictions as the voting mechanism for our above ensemble neural network architecture. Eq. 1 takes the average of each prediction made by all the neural networks, which are then passed through a hard threshold of 0.5 to classify manipulated and unmanipulated images.

The following section shows the extensive experimentation we have performed on every hyperparameter of our ensemble network and a comparison of accuracy for each.

3 Results and Experimentation

For our experiments, we commenced with k-fold cross-validation as a robust strategy for model evaluation. After carefully analyzing the dataset, we implemented a five-fold cross-validation on our ensemble neural network. Cross-validation is known to increase accuracy and decrease overfitting on smaller datasets[10]. In this approach, the dataset was diligently partitioned into five balanced folds. In each iteration, four folds were utilized for model training, while the remaining fold served as the test set. The model was trained for 1000 epochs, and the best-performing model's accuracy and F1 score were recorded. As illustrated in Fig. 5, we calculated mean accuracy and F1 scores to measure the overall performance of the models. Interestingly, the average F1 score and accuracy, which stood at 69.5% and 70%, respectively, were slightly below the original ensemble neural network performance metrics.

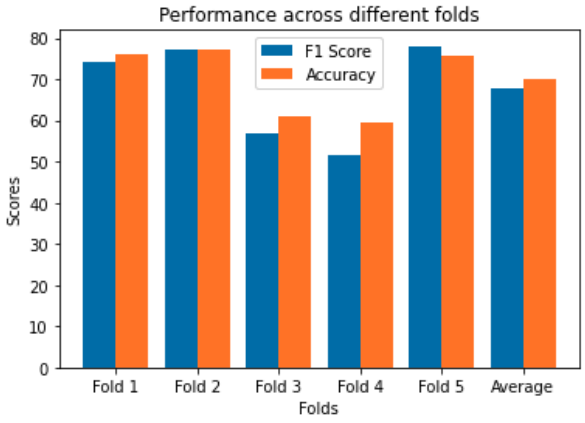

Fig. 5. Accuracy and F1 scores obtained from the 5-fold cross-validation of the ensemble neural network model

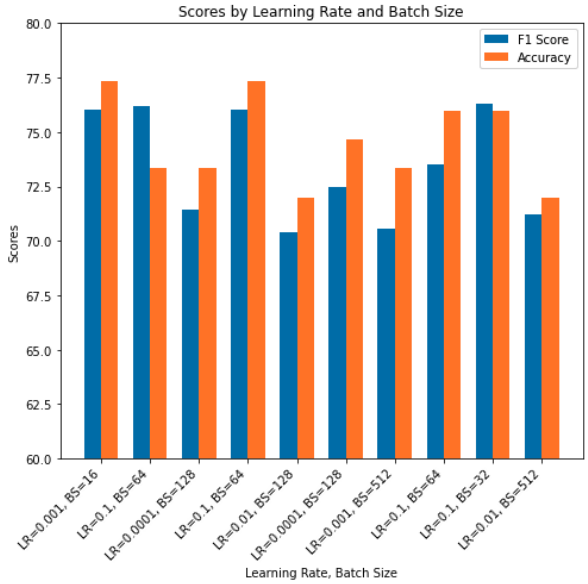

Fig. 6. Comparison of Model Performance using Different Hyperparameters. Graph illustrates the F1 scores and accuracies achieved by a model trained with various combinations of learning rates and batch sizes.

Next, we implemented hyperparameter tuning to enhance the model's accuracy. In this process, we tuned various parameters, including the number of neurons in the hidden layers, the learning rate, and the batch size of each neural network within our ensemble model. We employed the random search method for this purpose, known for its computational efficiency, aiming to optimize the parameters for the best accuracy. After experimenting

with different configurations, we selected four neural networks for our ensemble model. The architectures of these networks are as follows: [4, 128, 64, 1], [4, 256, 128, 32, 1], [4, 512, 128, 1], and [4, 128, 1]. Through this tuning process, we determined that a learning rate of (0.001, 0.1) and a batch size of (16, 64) yielded the best performance, achieving an accuracy of 77.33% and an F1 score of 76.05%, as illustrated in Fig.6.

Finally, we tuned the hard-coded classification threshold of 0.5. The choice of 0.5 as a threshold is conventional but not necessarily optimal [11]. We experimented with threshold values(θ) ranging from 0.4 to 0.7. The results, presented in Table 1, show variations in accuracy and F1 score corresponding to different thresholds. It is observed that as the threshold increases, the accuracy and F1 score generally tend to decrease. The highest accuracy is achieved at θ values of 0.40 and 0.50, while the F1 score peaks at 0.45 and 0.50. The confusion matrix reveals that a lower threshold is more effective in classifying manipulated images, while a higher threshold performs better with unmanipulated images.

Theta (θ)	Accuracy (%)	F1 Score (%)	TP	TN	FP	FN
0.40	76.0	74.6	29	28	8	10
0.45	74.6	75.9	29	27	8	11
0.50	76.0	77.3	30	27	7	11
0.55	73.3	71.4	30	25	7	13
0.60	73.3	70.5	31	24	6	14
0.65	72.0	68.6	31	23	6	15
0.70	73.3	67.7	34	21	3	17

Table 1. Performance Metrics and confusion matrix of our ensemble neural network at Various Thresholds (θ)

Using the methods described, we demonstrate that computers better detect manipulated images than humans. According to [4], the human mean accuracy in detecting manipulated images is 56%, and it was also found that humans are better at detecting unmanipulated images (61.3%) than correctly detecting manipulated images(50.1%). Comparing these results to our deep-learning techniques, it is evident that machines outperform humans in this task. Our ensemble neural network can accurately predict if the image is manipulated with an accuracy of 77.33% and an F1 score of 76.05%. Our neural network model achieved an accuracy of 76% and an F1 score of 73.5%. Meanwhile, our Decision tree model achieves an accuracy of 72% and an F1 score of 69.5%. Lastly, while being the least accurate among our models, our Maximum Likelihood model still achieves an accuracy of 69.3% and an F1 score of 63.5% [3]. Nevertheless, compared to human performance, all our models exceed in accuracy, as shown in Table 2.

Model	Accuracy (%)	F1 Score (%)
Humans	56	-
Maximum Likelihood	69.30	63.50
Decision Tree	72	69.50
Neural Network	76	73.50
Ensemble Neural Network	77.33	76.05
Ensemble Neural Network(k-fold)*	69.50	70.00
Ensemble Neural Network(random-search)**	77.33	76.05
Ensemble Neural Network(Threshold)***	76	74.60

Table 2. Comparison of accuracy and F1 scores for different models and human performance.

While accuracy and F1 score are essential metrics to evaluate model performance, we use another method called confusion matrices to provide in-depth insights into each model's performance. As detailed in Table 3, we can see that for the ensemble neural network model, out of 37 unmanipulated data points in the test set, it correctly classifies 30 data points, and out of 38 manipulated images, it correctly identifies 27. This indicates a balanced performance across the dataset. The neural network model, out of 37 unmanipulated data points, correctly identified 32 data points[3]. However, it is slightly less effective at detecting manipulated images than our ensemble neural network. While still surpassing human accuracy, the maximum likelihood technique is the least effective of the three, particularly when classifying manipulated images, suggesting that it struggles to identify manipulated data points[3].

Model	True Positive	True Negative	False Positive	False Negative
Ensemble Neural Network	30	27	7	11
Neural Network	32	25	5	13
Decision Tree	30	24	7	14
Maximum Likelihood	32	20	5	18

Table 3. Confusion matrix values for each model.

4 Discussion

Compared to the findings of [4], our results show the superior performance of machines in detecting manipulated images over humans. The grounded theory analysis conducted by [4] proved that participants often relied on logic, pre-existing beliefs, and knowledge when determining if an image was manipulated. However, in many cases, this logical decision proved counterproductive, resulting in the wrong classification of test images. Another finding was that the number of fixations and duration of eye gaze was directly proportional to accurately identifying the manipulated part of the image. Lastly, [8] reported similar findings, adding that participants with prior knowledge about the experiment and image manipulation techniques outperformed their other people. Our research is aligned with the above findings. Deep learning techniques tend to classify manipulated images better because of their data-driven nature. While humans often get sidetracked by irrelevant details in an image, disrupting their decision-making, machine learning techniques focus solely on the data they are trained on. Compared to humans, these models are devoid of presumptions, often leading to counterproductive results in human evaluations.

5 Conclusion and Future Work

Our research highlights the challenges of image manipulation detection tasks. We have implemented Deep learning techniques on the eye gaze dataset to classify manipulated images from the original. Our model outperforms humans, especially the ensemble neural network model, which shows a promising accuracy of 77.3%, significantly surpassing the human accuracy of 56%. Image editing software continues to advance rapidly, and Generative AI opens the door to uncharted territories in image editing. While this technology is astonishing, it also paves the way for increasingly sophisticated image manipulation techniques. So, we need to get better at detecting manipulated images. This research is our second stone in developing more robust image manipulation detection systems. We prove that more complex neural networks can better learn complex data and make more accurate predictions. Our extensive experimentation on the ensemble neural networks paved new paths of research in the field of deep learning.

Our future work involves working with other extensive datasets and building a more robust machine-learning model to accurately detect manipulated images and compare them to different state-of-the-art models. In the fast-paced world, there is a need for real-time manipulated image detection software. We are keen to partner with industry experts to add more features to our detection system and make our system accessible to the world.

In conclusion, we have just touched the surface level of image manipulation detection. Our RIMD2 performs better than our previous model, but we are optimistic that we can make it more robust with extensive research in the future. The journey ahead will surely be exciting and filled with challenges, and we are confident about building a more robust image manipulation detection system for the future.

References

1. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning, pp. 8821–8831. PMLR (2021)
2. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
3. Ahmed, J.: Towards Robust Image Manipulation Detection: A Comparative Study of Neural Networks and Statistical Methods vs Human Perceptions. (2023)
4. Caldwell, S., Gedeon, T., Jones, R., Copeland, L.: Imperfect understandings: a grounded theory and eye gaze investigation of human perceptions of manipulated and unmanipulated digital images. In: Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science, vol. 308, pp. 2 (2015)
5. Zhang, S., Liu, M., Yan, J.: The diversified ensemble neural network. Advances in Neural Information Processing Systems, vol. 33, pp. 16001–16011 (2020)
6. Itti, L., Koch, C.: A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention. Vision Research **40**, 1489-1506 (2000)

7. Harding, G., Bloj, M.: Real and Predicted Influence of Image Manipulations on Eye Movements During Scene Recognition. Journal of Vision **10**(2:8), 1-17 (2010)
8. Caldwell, S.B., Gedeon, T.D., Jones, R.L., Henschke, M.: Comparing eye gaze tracking to reported perceptions of manipulated and unmanipulated digital images. Aust. J. Intell. Inf. Process. Syst. **14**(3), 26–36 (2015)
9. Glaser, B.G., Holton, J.: Remodeling grounded theory. Historical Social Research/Historische Sozialforschung. Supplement, pp. 47–68. JSTOR (2007)
10. Berrar, D.: Cross-Validation. (2019)
11. Milne, L., Gedeon, T., Skidmore, A.: Classifying dry sclerophyll forest from augmented satellite data: Comparing neural network, decision tree and maximum likelihood. In: ACNN'95. (1995)
12. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001). Springer.
13. Richards, J.: Remote Sensing Digital Image Analysis. 2nd edn. Springer Verlag (1993)