Improving the Accuracy of Identifying Real-Time Indian Twins Using CNN Compared with Random Forest

Vallipi Dasaratha* and J. Joselin Jeya Sheela†
Department of Electronics and Communication Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, 602105, India

Keywords: Biometric, Convolutional Neural Network, Face Recognition, Identification, Image Processing, Novel, Technology.

Abstract: The objective of this study is to achieve real-time identification and analysis of Indian twins using the random forest algorithm, while also comparing its performance with the Convolutional Neural Network (CNN) algorithm in terms of accuracy. Materials and Methods: For the purpose of face recognition of twins with face and ID recognition, the random forest algorithm is chosen over the Convolutional Neural Network (CNN). The study involves two groups, namely Group 1 and Group 2, with an overall sample size of 1430 and 20 sample iterations for each group. Results and Discussion: The comparison and classification of real-time Indian twins are conducted using the Random Forest algorithm and the performance is measured using the CNN algorithm. The achieved accuracy rates are 52.3965% for Random Forest and 64.305% for CNN. By comparing the accuracy of both algorithms, independent sample tests reveal a statistically significant difference with a p-value of 0.001 ($p<0.05$), confirming the significance of the hypothesis through an independent sample t-test. Conclusion: This study evaluated the effectiveness of two image processing algorithms, namely Random Forest and CNN. The results indicate that Random Forest achieves an accuracy of 52.3965%, outperforming CNN which achieved an accuracy of 64.3050%. This suggests that for identification using ID recognition, Random Forest provides superior performance compared to CNN.

1 INTRODUCTION

There are diverse applications for this intriguing challenge, including social media photo tagging, activity tracking, crime detection, and more. It poses a intricate visual challenge (FGVC) due to the minute inter-class variances of the twin objects. Owing to their remarkable accuracy, they are frequently employed (Chandana.S, Harini, and Senthil 2022). The utilization of object detectors like Yolo3, Faster-RCNN, SSD, and ResNet-101, alongside pretrained base networks such as VGGNet, ResNet-101, Inception with ResNet, and Retina Net, has shown promising outcomes. Nevertheless, the accuracy of all these models diminishes when objects with exceedingly slight differences need to be recognized. CNNs acquire significant low-level features in a hierarchical, feed-forward manner, which may lead to smoother learning progression (Chandana.S, Harini, and Senthil 2022; Lane et al. 2015), impacting the model's adeptness in detecting fine-grained objects. The incorporation of features taught at diverse levels and scales is essential for this objective. To avert overfitting, a substantial dataset is required. We devised a solution by employing publicly available images of renowned twins Mary-Kate and Ashley Olsen, generating an annotated dataset of 120 images to address my twin identification challenge (Shoji and Zhang 2019). This dataset comprises images of the twins standing alongside other individuals, singly or together, constituting my dataset's principal challenge. To surmount this, I employed a VGGNet CNN trained with the ImageNet dataset, and applied the Single Shot Detector (SSD) based on the trained VGGNet base network. SSD is a rapid one-pass detector (B402. et al. 2021) with low computational demands suitable for video detection. Despite a marginal trade-off in accuracy when objects are very small, SSD excels in recognizing them by capturing features across diverse scales. This study's primary

* Research Scholar
† Research Guide, Corresponding Author

Dasaratha, V. and Sheela, J.
Improving the Accuracy of Identifying Real-Time Indian Twins Using CNN Compared with Random Forest.
DOI: 10.5220/0012520100003739

In *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics (AI4IoT 2023)*, pages 487-493
ISBN: 978-989-758-661-3

contributions encompass photographing the "Olsen" twins in their natural environment to comprehend twin identification complexities and providing an effective method for fine-grained item detection in scenarios with limited datasets. Detailed exploration of current fine-grained object detection research is covered in the second part of this article (Gharbi et al. 2018; Hegde and Manjunatha 2022). The third and fourth sections describe the dataset and model, respectively, while Section 5 presents the research findings.

In my scenario, there exist exceedingly subtle distinctions between the two distinct groups. My twin subjects, referred to as "Mary" and "Ashley," represent the classes. These classes exhibit extremely delicate dissimilarities that render their differentiation challenging across all images (Mahapatra, S et al. 2016). The novel learning model must diligently capture these distinguishing attributes to the maximum extent possible. Moreover, it should not hastily incorporate irrelevant features, as this would obscure the significance of discriminative characteristics in the detection process. Employing transfer learning in CNN models expedites training while augmenting novel learning, particularly when the dataset is constrained (Vinod and Padmapriya 2022). Through pre-training a model on a relevant dataset, it can be adapted to a different task using a smaller dataset for that purpose. Despite employing distinct datasets for the two tasks, the model's weights can be adjusted. Employing a slightly different approach, I extended this notion to address the twins' detection challenge. My dataset encompasses images with and without immediately distinguishable core attributes. Following the initial update of my pre-trained model with the non-essential images, only the final layer was modified using the critical subset. This approach intuitively aligns with CNN learning, allowing the utilization of the most distinctive attributes within the upper layers (DeVerse and Maus 2016).

2 MATERIALS AND METHODS

The experimentation took place within the Machine Learning Laboratory at Saveetha School of Engineering, Saveetha Institute of Medical and Technological Sciences. Two distinct entities were established for the study: Group 1 denoting CNN and Group 2 corresponding to Random Forest. Employing a G-power of 80%, the system calculates the required sample size and defines it as 40 iteration samples while accessing the Clincale website (Group 1 - 20, Group 2 - 20). The setup consists of two separate groups, with a cumulative sample size of 1430. Each of the two groups, Group 1 and Group 2, underwent 20 sample iterations.

The sample preparation process for the renowned Random Forest machine learning technique within the context of the supervised technology learning novel approach has been completed. This approach holds potential advantages for machine learning tasks encompassing both classification and regression. It is rooted in the innovative ensemble learning theory, a strategy that integrates multiple classifiers to address challenging problems and enhance the overall model performance.

Sample preparation for Group 2, focused on CNN, has been completed. CNN, as a classification algorithm, accomplishes the task by transforming the initial training data into a higher-dimensional space through a nonlinear mapping. Within this transformed space, Random Forest seeks an optimal hyperplane that effectively separates instances of distinct classes. The determination of this hyperplane is influenced by the cases that closely interact with the division between the two classes, known as support vectors. The effectiveness of the Random Forest classifier is significantly influenced by the chosen kernel function and its associated parameters.

The most favorable outcomes are often achieved by employing a Puk kernel along with a kernel parameter value of C = 1.

The development of the face recognition identification system was carried out using Jupyter Notebook on a Windows 11 operating system. The system's implementation involves two distinct groups: Random Forest and CNN methods. These algorithms are integrated into a novel dataset, which is then subjected to training and testing processes to enhance accuracy. The sample dataset consists of 40 instances. During the model training, various loss functions were employed. To better align with the correct labels, the initial cross-entropy loss was adjusted to focal loss, as inspired by the SSD study investigation.

The cross-entropy loss balances the weights of both positive and negative instances, but it doesn't differentiate between simple and complex cases. In contrast, focal loss reshapes the cross-entropy loss by reducing the weight applied to well-classified or simple data. The focal loss function is explained for classification, with "alpha" representing the balancing parameter and "gamma" representing the focusing parameter.

3 RANDOM FOREST

The widely recognized Random Forest is a machine learning algorithm that employs an innovative ensemble learning technique for prediction purposes. It can be employed in supervised learning scenarios for both classification and regression tasks. The algorithm's functionality involves training numerous individual classifiers, referred to as decision trees, and subsequently amalgamating their predictions to formulate a final prediction. This approach proves efficacious in enhancing the model's performance and addressing intricate challenges.

Algorithm for random forest

Step1:from sklearn.datasets import fetch_olivetti_faces

Step2: from sklearn.model_selection import train_test_split

Step3: from sklearn.metrics import accuracy_score

Step4: from sklearn.ensemble import RandomForestClassifier

Step5:X-data.reshape((data.shape[0],data.shape[1]*data.shape[2]))

Step6:X_train, X_test, y_train, y_test=train_testsplit(X,target, test_size=0.3, stratify=target, random_state=0)

Step7:clf = RandomForestClassifier()

Step8:clf.fit(X train pca, y train)

Step9:y pred-clf.predict(X_test_pca)

Step10: print(accuracy_score(y_pred,y_test)* 100)

3.1 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs), often known as convnets or CNNs, constitute a crucial facet of machine learning. They represent a subset of artificial neural network architectures that find application in diverse and pioneering objectives and datasets. Specifically tailored for deep learning algorithms, CNNs serve as a network architecture designed for image recognition and the processing of pixel data.

4 STATISTICAL ANALYSIS

The investigation was conducted using IBM SPSS version 21. The research independent factors encompassed the project, V name, and year end. Meanwhile, the research dependent variables involved face and face ID. Iterations with a limit of 15 instances were executed for both the proposed and existing methods. The anticipated accuracy for each innovative iteration was recorded for accuracy analysis. Subsequently, the results of these iterations were subjected to an Independent Sample T-test. A p-value of $p=0.350$ ($p<0.05$) indicated that there was no discernible disparity in the accuracy of the algorithms ("Real-Time Modeling of Albedo Pressure on Spacecraft and Applications for Improving Trajectory Est." 2023).

5 DATASET

A dataset of widely available photos featuring the Olsen twins was meticulously assembled. Employing a PythonScript, images were scraped from Google Images, subsequently annotated using the Labeling tool. In total, 120 photos were amassed, out of which 50 were allocated for testing purposes, and 800 were designated for training and validation. The classes were denoted as "mary" and "ashley." Annotating the images presented a challenge due to the fact that numerous photographs lacked clear facial distinctions between the twins.

6 RESULTS

Table 1 clearly indicates that Random Forest outperformed CNN significantly in the context of identification using face and ID recognition. The precision and performance of Random Forest surpassed those of CNN, signifying its superiority for this specific dataset and task-

Table 2 presents the efficacy of CNN and Random Forest on a face and ID recognition dataset. The outcomes demonstrate that Random Forest achieved a mean accuracy of 52.3965, accompanied by a standard deviation of 1.49933 and a standard error mean of 0.33526. In contrast, CNN exhibited a mean accuracy of 64.3050, with a standard deviation of 1.34707 and a standard error mean of 0.30121.

Table 1: Comparison between Random Forest and CNN with N=20 iteration samples of the dataset with the highest accuracy 72% and 66% respectively in the sample (when N=1) using the dataset size=7476 and the 66.45% of training & 62.30% of testing data.

Sample (N)	Dataset Size / rows	Random Forest accuracy in %	CNN accuracy in %
1	7182	72.15	66.66
2	7123	72.10	66.45
3	6987	72.05	66.10
4	6900	71.98	65.89
5	5087	71.87	65.50
6	5012	71.77	65.28
7	4987	71.56	65.18
8	4565	71.45	64.50
9	4444	71.18	64.38
10	4321	70.67	64.28
11	4312	70.48	63.96
12	4300	70.34	63.86
13	3099	70.28	63.67
14	3081	70.16	63.54
15	3097	69.76	63.23
16	3000	69.46	63.78
17	2098	68.89	63.78
18	2012	68.47	62.72
19	1089	67.66	62.60
20	1001	66.45	62.30

Table 2: Statistical result of Random Forest algorithm and CNN algorithm. Mean error value, SD and standard error mean for RF and CNN algorithm are obtained for 20 iterations. It is observed that the mean for Random Forest (52.3965%) performed better than the CNN (64.305%) algorithm.

Group Statistic					
	ALGORITHMS	N	Mean	Std. Deviation	Std. Error Mean
ACCURACY	Random Forest	20	52.3965	1.49933	.33526
	CNN	20	64.3050	1.34707	.30121

In Table 3, the results of the significance test reveal a substantial distinction in the accuracy of the two algorithms. The significance value of less than $p=0.350$ ($p<0.05$) underscores the preference for CNN over Random Forest for this dataset and task.

Figure 1 graphically portrays the mean accuracy of identification using face and ID recognition for both Random Forest and CNN. The depicted results underscore that Random Forest attained an accuracy of 52.3965%, whereas CNN achieved an accuracy of 64.3050%. This underlines CNN's better performance compared to Random Forest on this dataset and task.

Table 3: The independent sample t-test of the significance level Random Forest and CNN algorithms results with significant values ($p < 0.05$). Therefore, both the Random Forest and the CNN algorithms have a significance level less than 0.02 with a 95 % confidence interval.

Independent samples test									
Accuracy	Levene's Test for Equality of Variances		T-test of Equality of Means					95% of the confidence interval of the Difference	
	F	Sig.	t	df	Sig (2-tailed)	Mean Differ-ence	Std Error Differen-ce		
Equal Variance Assumed	.103	.350	13.699	38	0.01	6.17400	.45070	5.28181	7.08639
Equal Variance Not Assumed			13.699	37.572	0.01	6.17400	.45070	5.28127	7.08673

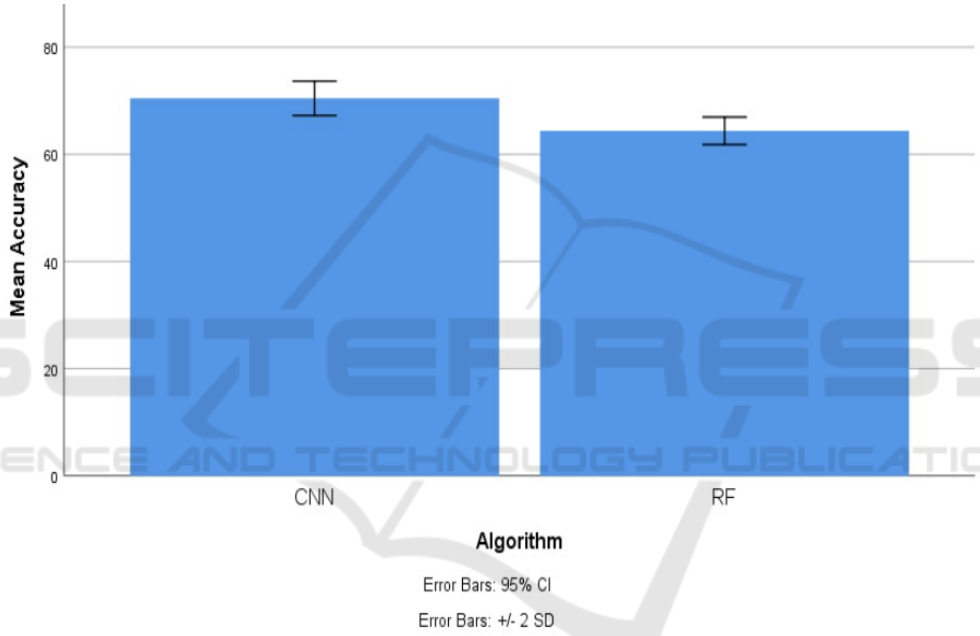

Figure 1: Comparison of precision between the CNN algorithm and RF The mean precision of the CNN algorithm is better than the RF, and the standard deviation of the CNN algorithm is highly better than the RF. X-axis: CNN algorithm vs RF Algorithm and Y-axis represents Mean Precision values ±2 SD.

7 DISCUSSION

The aforementioned research study demonstrated that Random Forest achieved a higher accuracy of 52.3965% as compared to CNN's accuracy of 64.3050%. A statistically significant difference between the accuracy of the two algorithms was determined through the utilization of independent sample t-tests, yielding a p-value of $p=0.350$ ($p<0.05$) (Nasri and Kargahi 2012). The dataset employed for this study was sourced from the open-source platform Kaggle and was applied for identification through face and ID detection. In the current system, Random Forest displayed superior accuracy to CNN, achieving 52.3965% accuracy in contrast to CNN's 64.3050%. For the proposed system, the dataset was trained and tested using applications such as SPSS and Jupyter Notebook, which were also used for forecasting graphs (DeVerse and Maus 2016).

In the proposed system for fake voter identification using face and ID recognition, it is anticipated that Random Forest will exhibit higher accuracy compared to CNN. The performance of various classifiers including Random Forest, KNN,

CNN, etc., is evaluated using an independent dataset, a task that presents challenges due to limited available data (Yadav and Kumar, n.d.; Gao, Xu, and Wang 2003). Assessing classifier performance is intricate when comparing different learning methods, as it involves evaluating the error rate, which determines the classifier's success in correctly categorizing instances (Agarwal et al. 2020). This evaluation is achieved by considering the mistakes made by the classifier in each instance. To effectively gauge classifier performance, independent test data not used in the model is employed. If additional data is required, it can be partitioned into training and testing sets.

Increasing the volume of training data leads to higher classification accuracy and enhances the utility of testing data. Nevertheless, a challenge emerges when the available data is insufficient. To address this, manual separation of training and test data is essential (Samek et al. 2019) (Ramkumar, G. et al. 2021). Insufficient data can also introduce issues. To mitigate this, the holdout approach is commonly employed, allocating one-third of the data for testing and the remaining two-thirds for analysis. Cross-validation is another effective strategy, necessitating a decision on the number of data folds or partitions to utilize. In this research, a 10-fold cross-validation method was adopted, splitting the data into ten segments with equal representation across classes (Gunjan and Zurada 2020). This approach involves dividing the data into ten equal parts and iteratively using 10% for testing and 90% for training. After each iteration, one tenth is designated for testing. This process allows for estimating the overall error over ten iterations (Khanna et al. 2021).

A notable limitation of the twin study research method lies in the potential influence of significant gene-environment correlations or interactions. Such factors can introduce inaccuracies when attempting to segregate liability into distinct genetic and environmental components. In the realm of technology, a parallel concept to twins emerges through the utilization of Indian twin technology. This concept, prevalent within the industrial sector, involves creating digital replicas of objects or processes. To achieve this, sensors are strategically positioned to collect real-time data from the physical process, which is then fed into AI systems for processing. Subsequently, these digital twins offer a platform to comprehensively examine and simulate the operational mechanics of the object or process, facilitating in-depth insights into product behavior and performance simulations.

8 CONCLUSION

The research study focused on the evaluation of two image processing algorithms, Random Forest and CNN, for the purpose of identification using face and ID recognition. The results revealed that Random Forest exhibited a higher accuracy of 52.3965% in contrast to CNN's accuracy of 64.3050%. These findings signify that Random Forest outperforms CNN in the realm of ID recognition-based identification.

REFERENCES

Agarwal, Basant, Valentina Emilia Balas, Lakhmi C. Jain, Ramesh Chandra Poonia, and Manisha Sharma. (2020). Deep Learning Techniques for Biomedical and Health Informatics. Academic Press.

B402., Escuela Politécnica Superior Uam, Escuela Politécnica Superior Uam, B402., and Hadi Abooei Mehrizi. (2021). “Identifies Polyps in Real Time with Accuracy 96.67% in Screening Colonoscopy Using Convolutional Neural Networks (CNN).” IBJ Plus.

Chandana.S, Harini, Chandana S. Harini, and Kumar R. Senthil. (2022). “A Deep Learning Model to Identify Twins and Look Alike Identification Using Convolutional Neural Network (CNN) and to Compare the Accuracy with SVM Approach.” ECS Transactions. https://doi.org/10.1149/10701.14109ecst.

DeVerse, Shawn, and Stefan Maus. (2016). “Improving Directional Survey Accuracy through Real-Time Operating Centers.” Day 2 Tue, August 23, 2016. https://doi.org/10.2118/180652-ms.

Gao, Jingbo, Minqiang Xu, and Rixin Wang. (2003). “Study About Real-Time Finite Element Method Using CNN.” Computer Technology and Applications. https://doi.org/10.1115/pvp2003-1908.

Gharbi, Salem Al, Salem Al Gharbi, Qinzhuo Liao, Salaheldin Elkatatny, and Abdulazeez Abdulraheem. (2018). “IIncreasing ANN Accuracy, by Improving the Training Dataset Criteria. Case Study: Identify the Formation Density from The Drilling Surface Parameters in Real-Time.” All Days. https://doi.org/10.2118/192363-ms.

Gunjan, Vinit Kumar, and Jacek M. Zurada. (2020). Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2020. Springer Nature.

Hegde, Chandan R., and H. T. Manjunatha. (2022). “Real Time Indian Traffic Sign Detection Using Image Processing and CNN.” International Journal of Advanced Research in Science, Communication and Technology. https://doi.org/10.48175/ijarsct-5363.

Khanna, Ashish, Deepak Gupta, Zdzisław Pólkowski, Siddhartha Bhattacharyya, and Oscar Castillo. (2021). Data Analytics and Management: Proceedings of ICDAM. Springer Nature.

Lane, D., S. Hill, J. Huntingford, P. Lafuente, R. Wall, and K. Jones. (2015). “Effectiveness of Slow Motion Video Compared to Real Time Video in Improving the Accuracy and Consistency of Subjective Gait Analysis in Dogs.” Open Veterinary Journal.

Mahapatra, S., Vickram, A. S., Sridharan, T. B., Parameswari, R., & Pathy, M. R. (2016). Screening, production, optimization and characterization of β-glucosidase using microbes from shellfish waste. 3 Biotech, 6, 1-10.

Nasri, Mitra, and Mehdi Kargahi. (2012). “A Method for Improving Delay-Sensitive Accuracy in Real-Time Embedded Systems.” 2012 IEEE International Conference on Embedded and Real-Time Computing Systems and Applications. https://doi.org/10.1109/rtcsa.2012.39.

Ramalakshmi, M., & Vidhyalakshmi, S. (2021). GRS bridge abutments under cyclic lateral push. Materials Today: Proceedings, 43, 1089-1092.

Paramasivam, G., Palem, V. V., Sundaram, T., Sundaram, V., Kishore, S. C., & Bellucci, S. (2021). Nanomaterials: Synthesis and applications in theranostics. Nanomaterials, 11(12), 3228.

Ramkumar, G. et al. (2021). “An Unconventional Approach for Analyzing the Mechanical Properties of Natural Fiber Composite Using Convolutional Neural Network” Advances in Materials Science and Engineering vol. 2021, Article ID 5450935, 15 pages, 2021. https://doi.org/10.1155/2021/5450935

(“Real-Time Modeling of Albedo Pressure on Spacecraft and Applications for Improving Trajectory Est.” (2023). https://doi.org/10.2514/6.2023-2207.vid.

Samek, Wojciech, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. (2019). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer Nature.

Shoji, Yuta, and Lifeng Zhang. (2019). “Research for Improving Identification Accuracy of Specific Fish Species with CNN.” Proceedings of The 7th International Conference on Intelligent Systems and Image Processing 2019. https://doi.org/10.12792/icisip2019.036.

Vinod, G., and G. Padmapriya.(2022). “An Adaptable Real-Time Object Detection for Traffic Surveillance Using R-CNN over CNN with Improved Accuracy.” 2022 International Conference on Business Analytics for Technology and Security (ICBATS). https://doi.org/10.1109/icbats54253.2022.9759030.

Yadav, Anu, and Ela Kumar. n.d. “Instance Segmentation for Real-Time Video Detection Using FPN and Mask R-CNN.” Zohourian, Farnoush, Borislav Antic, Jan Siegemund, Mirko Meuter, and Jose

Pauli. (2018). “Superpixel-Based Road Segmentation for Real-Time Systems Using CNN.” Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications.