Transfer Learning Technique for Deepfake Face Detection Using Weighted Average Ensemble Model

Jannatul Mawa[1] and Md. Humayun Kabir[2]
[1,2]Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh
Correspondence E-mail: hkabir@juniv.edu

***Abstract*—Deepfakes represent a significant cybersecurity threat with their ability to create highly convincing fraudulent media. As deepfake technology becomes more sophisticated and accessible, the potential for cybercrimes such as identity theft, fraudulent account openings, and financial scams increases. To address the rising threat of deepfakes, this research explores detecting deepfake face images by combining transfer learning with an ensemble technique. Four pre-trained models have been employed for the transfer learning task. Finally top three performing models were combined for the ensemble. The ensemble model has been evaluated against a benchmark dataset, namely 140K Real and Fake Faces. The ensemble model significantly surpassed the individual models, achieving an accuracy of 81.25%. This research demonstrates the potential of integrating multiple pre-trained models to improve deepfake image detection, laying a strong foundation for future advancements.**

***Index Terms*— Fraudulent media, Identity theft, Deepfake face images, Ensemble model, Pre-trained models, Deepfake image detection.**

I. INTRODUCTION

The most identifiable aspect of a person is their face. The growing advancement of facial synthesis technology has made the security risks associated with face modification more critical [1]. With more than a billion images uploaded daily to Instagram and vast quantities of selfies stored by Google, almost everyone now has a digital presence. These digital footprints, from LinkedIn photos to family pictures on Facebook, can be exploited by AI to produce convincing deepfakes for malicious uses [2]. Deepfakes are an emerging threat in cybersecurity, leveraging AI to create convincing false media that can be used for various malicious purposes, including harassment, blackmailing and misleading contents [3]. With the proliferation of digital images and personal data online, the risk of deepfake-enabled cybercrimes is increasing.

In this work we tried to investigate how existing pretrained models can contribute to the detection of deepfakes. How accumulating the predictions from multiple models significantly improves the overall detection accuracy. To assess the effectiveness of the model against benchmark deepfake dataset. To contribute innovative methodologies to the field through the effective use of transfer learning and ensemble techniques.

The subsequent sections are structured as follows: section II describes the related work in this area. Section III provides background study on the key tools and technologies essential to conduct research in the specified research problem. Section IV provides an overview of the sample data set used in this research work to complete the thesis work. Section V presents the system model describing the step-by-step deepfake face detection procedure.

II. RELATED WORK

This section navigates through current research to uncover pivotal insights and methodological approaches. Fernando et al. [4] employed adversarial training techniques and subsequently used attention-based mechanisms to detect hidden facial manipulations. Luca et al. [5] focused on identifying and extracting fingerprints indicative of convolution traces from the GAN image generation process, utilizing the Expectation-Maximization algorithm for detection. Li et al. [6] proposed using a face X-ray technique to detect traces of modification around the boundary region of an artificial face. Xu et al. [7] introduced an approach to detect Deepfake videos by creating texture features and applying a feature selection technique. The discriminative feature vector derived from this process is subsequently utilized for classification using SVM. Zhang et al. [8] developed a GAN simulator that mimics common artifacts in GAN-generated images and uses these artifacts as input for a classifier to detect deepfakes. Tariq et al. [9] suggested employing neural networks for identifying deceptive GAN videos. Their approach involves analyzing statistical elements of images to improve the detection of artificially generated fake facial photographs. Ismail et al. [10] proposed a new method for deepfake detection utilizing Extreme Gradient Boosting. They used the YOLO detector to isolate the face region from video frames. Then, features from these faces were extracted using InceptionResNetV2. These features were then fed into XGBoost, which acted as a recognition system built on top of the CNN architecture. Xuan et al. [11] utilized image preprocessing methods, including Gaussian blur and Gaussian noise. These techniques enhance the mathematical resemblance between genuine photographs and counterfeits at the pixel level. As a result, the scientific classifier can capture more inherent features, thereby improving its generalization capability compared to previous techniques in image forensics. Wang et al. [12] showed that by employing meticulous pre-processing, post-processing, and data augmentation techniques, a conventional classifier trained on ProGAN—an unconditional CNN generator—can exhibit remarkable generalization capabilities across unfamiliar architectures, datasets, and training methodologies.

III. BACKGROUND

This section offers a concise overview of the models and techniques employed in this study, setting the stage for a

deeper exploration into their application within the context of transfer learning and weighted average ensemble methods.

A. Transfer Learning

Transfer learning is a technique in machine learning that leverages knowledge acquired from one task to enhance learning performance on a related task. This method capitalizes on the knowledge gained from a large dataset, reducing training time and improving performance on new tasks [13]. The pre-trained models extract relevant features from new data and a new classifier trained on these features. This technique continues to enhance AI capabilities, providing efficient and effective solutions across diverse fields. Brief description of the models used in this study are as follows:

1) VGG16

The VGG16 [14] model, is a widely recognized deep learning model known for its simplicity and depth. This neural network consists of 16 weight layers. One of the defining features of VGG16 is its application of small (3x3) convolutional filters, which allows it to grasp intricate features while maintaining computational efficiency. Despite its relatively simple architecture, VGG16 has proven to be highly effective in various image classification tasks.

2) ResNet-50

ResNet50 [15], a part of the ResNet (Residual Networks) family, revolutionized deep learning by introducing residual learning. This model comprises 50 layers. The residual block of the model allows the network to learn residual functions. This architecture allows for the training of significantly deeper networks without a decline in performance, leading to significant improvements in accuracy.

3) InceptionV3

InceptionV3 [16], developed by Google, is part of the Inception series of models that aim to optimize both depth and width of the network. InceptionV3 introduces several enhancements, including factorized convolutions and aggressive regularization, which improve both the efficiency and accuracy of the model. This versatility makes InceptionV3 a powerful tool for a wide range of computer vision tasks.

4) Densenet-201

DenseNet201 [17], part of the Dense Convolutional Network (DenseNet) family, introduces a novel connectivity pattern where every layer is linked directly to every other layer in a feed-forward fashion. This results in 201 layers, where each subsequent layer receives input from the feature maps of all earlier layers. This architecture is particularly advantageous for tasks requiring detailed feature extraction and robust learning.

B. Weighted Average Ensemble

The Weighted Average Ensemble approach combines the strengths of multiple models to enhance overall performance [18]. In this method, predictions from different models are weighted and averaged to produce a final prediction. The weights are typically determined based on the individual performance of each model on a test set. This ensemble method leverages the complementary strengths of different architectures, leading to improved accuracy and robustness compared to any single model. Weighted Average Ensembles are particularly useful in applications where maximizing performance is crucial.

$$P_{ensemble} = \sum_{i=1}^{n} w_i * p_i \quad (1)$$

C. StyleGAN

StyleGAN [19], developed by NVIDIA researchers, introduces a style-based generator that enables precise control over image features at different levels. This allows for modifications in aspects like facial expressions, hair style, and even background details. This model incorporates style mixing, where the latent space vector is mapped into different styles applied at various stages of the image synthesis process. This helps generate more diverse and realistic images. Its applications span across creative industries and research, though it also raises important ethical considerations, particularly regarding potential misuse in creating deepfakes and other deceptive content.

IV. DATASET OVERVIEW

The "140K Real and Fake Faces" dataset hosted on Kaggle [20] has been employed in this study. This dataset includes a total of 140,000 images, with an equal distribution of real and fake faces. The real images in this dataset are sourced from NVIDIA's Flickr dataset [21] and the fake images are generated by StyleGAN [19]. Fig. 1 illustrates several images from the dataset.

Fig. 1. Real and fake images from the dataset [1].

V. METHODOLOGY

This section presents a comprehensive overview of the development and evaluation of an ensemble neural network designed for deepfake detection. It covers various aspects including the data preprocessing steps, transfer learning setup, ensemble model construction and the training methodologies employed.

A. System Architecture

The following Fig.2 illustrates the workflow of deepfake face detection process using an ensemble approach.

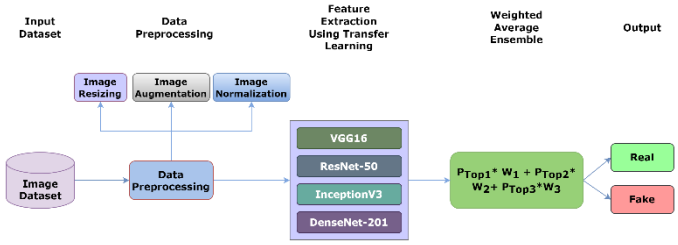

Fig. 2. Workflow of the deepfake face detection procedure.

B. Data Preparation

From the extensive collection of 140000 images, 2992 images were selected as the representative subset for the study. Table I demonstrates the distribution of images in the dataset.

Table I. Distribution of images

Class	Training	Validation	Testing	Σ
Real	1000	248	248	1496
Fake	1000	248	248	1496
Σ	2000	496	496	2992

All the images in the dataset were resized to a uniform resolution of 224 x 224. This step ensures consistency across the models in the ensemble. Pixel values of the images were normalized to the range [0, 1] by dividing each pixel value by 255. This normalization aids in optimizing the training convergence of neural networks.

C. Transfer Learning Setup

The deepfake detection system was constructed by leveraging an ensemble of pre-trained models, specifically VGG16, ResNet50, InceptionV3 and DenseNet201, each fine-tuned for the task of identifying fake images. Initially, these models were loaded with weights trained on ImageNet [22]. To adapt them for binary classification (real vs. fake images), the top layers were customized by incorporating a Global Average Pooling layer to reduce dimensionality, a Dropout layer to prevent overfitting, a Dense layer with 512 units and ReLU [23] activation for non-linearity, and an output Dense layer with one unit and Sigmoid [24] activation for binary classification. During the fine-tuning process, only the newly added layers were trainable, while the convolutional base layers were kept frozen to maintain the learned features from previous training.

D. Training Process

To optimize the performance of models in detecting deepfake images, a comprehensive training process was employed. Data augmentation techniques such as horizontal flipping was used to enhance dataset variability. The binary cross-entropy [25] loss function was employed to measure the discrepancy between predicted and actual labels, while the Adam [26] optimizer was employed to minimize the loss function and update model parameters. Each model underwent training for 20 epochs. To prevent overfitting, dropout rate of 0.2 was applied. Batch size was set to 64 balance training efficiency.

E. Ensemble Model Construction

The ensemble model was developed using a custom Weighted Average Layer, where the predictions of individual models were combined with manually assigned weights to generate a final prediction. From the four pre-trained models, the top three models—VGG16, DenseNet201, and InceptionV3 were selected based on their performance on the testing set. The weights were assigned as follows: VGG16 (0.3), DenseNet201 (0.5), and InceptionV3 (0.2). The pre-trained and fine-tuned models, saved in `.h5` format, were loaded to construct the ensemble. The ensemble classifier was further trained for an additional 10 epochs to fine-tune the combination of model predictions.

VI. RESULTS

This section presents the results of the experiments conducted to evaluate the deepfake image detection system. It includes the performance metrics, accuracy and loss scores, confusion matrices and prediction result.

A. Performance Metrics

Table II presents the performance of each individual models as well as the ensemble model across the dataset.

Table II. Performance metric summary for individual and ensemble model.

Model	Accuracy	Precision	Recall	F1-Score
VGG16	0.75	0.75	0.75	0.75
ResNet-50	0.58	0.58	0.58	0.58
InceptionV3	0.73	0.73	0.73	0.73
Densenet-201	0.79	0.80	0.79	0.79
Weighted Avg. Ensemble	0.81	0.81	0.81	0.81

ResNet-50's performance is the lowest, with all the scores at 0.58, highlighting its limitations. VGG16 is more reliable, consistently scoring 0.75 across all metrics. InceptionV3 is slightly behind VGG16, with all metrics at 0.73. DenseNet-201 shows strong performance, achieving an accuracy and F1-score of 0.79, and a precision of 0.8. This indicates a well-balanced model. The Weighted Average Ensemble model outperforms all others, with the highest score of 0.81 at all metrics, demonstrating its exceptional capability and robustness.

B. *Performance Evaluation*

Fig. 3 provides the performance results for five models: ResNet-50, VGG16, InceptionV3, Densenet-201, and the Weighted Average Ensemble.

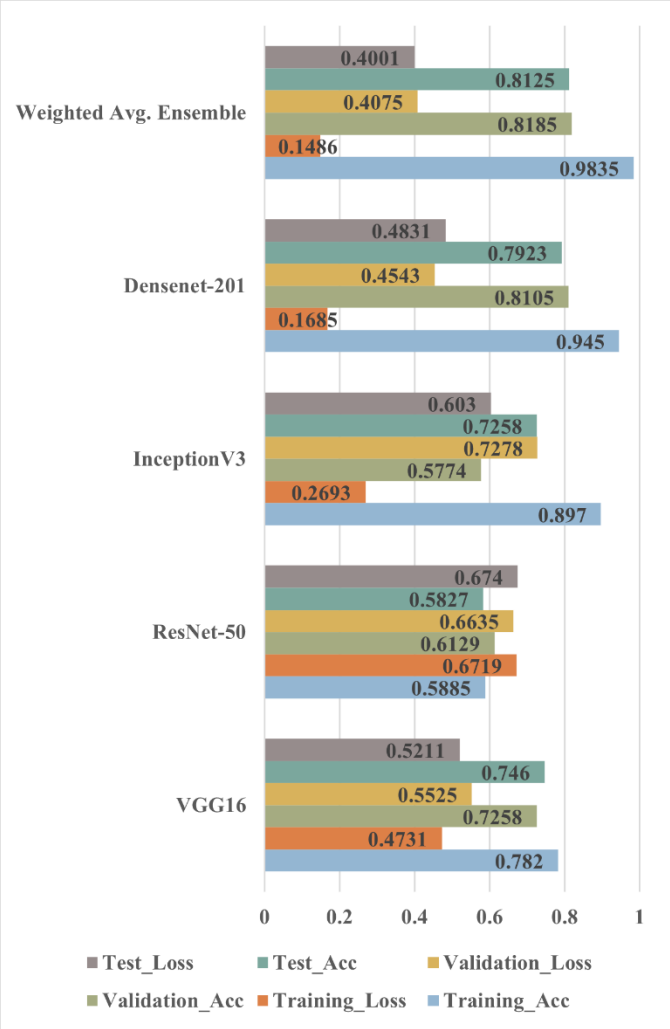

Fig. 3. Accuracy and Loss scores of the models.

VGG16 achieves high accuracy and generalizes well. ResNet50's validation accuracy exceeds its test accuracy, indicating overfitting. InceptionV3 shows very high training accuracy with some overfitting, but performs well on the test set. DenseNet201 displays the highest training accuracy and strong validation and test performance, suggesting good generalization. The ensemble model exhibits exceptional performance with the highest accuracy (0.8125) and lowest loss, indicating excellent generalization. It outperforms the other models, followed closely by DenseNet201 and VGG16, both of which show high training and test accuracies.

C. *Confusion Matrices*

Fig. 4 displays the confusion matrices for individual models.

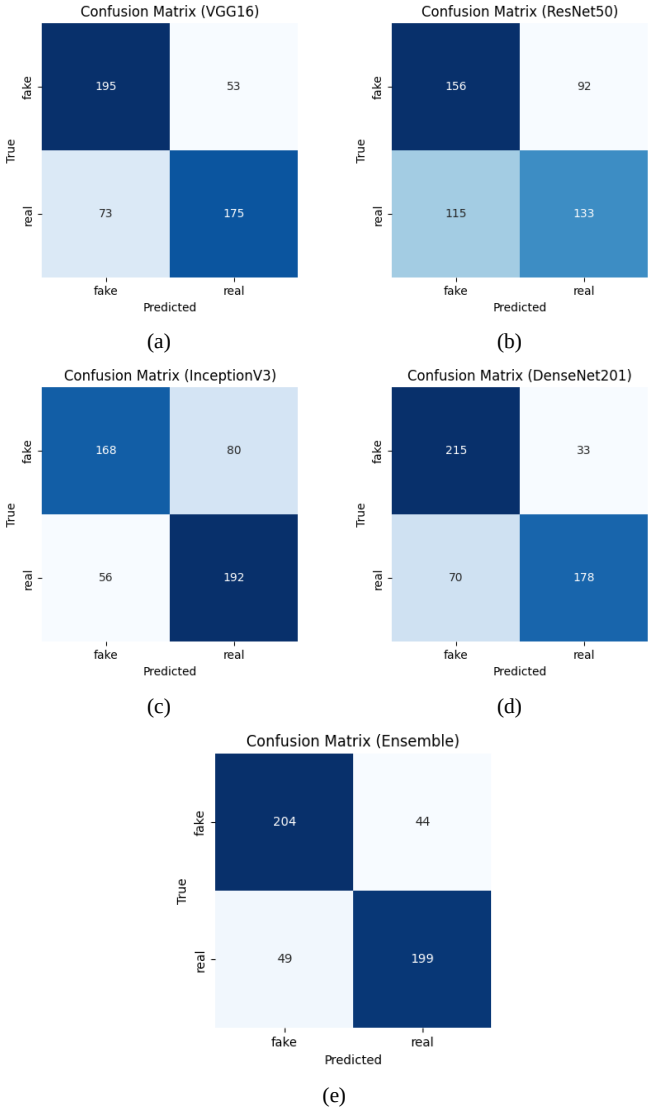

Fig. 4. Confusion Matrices for all the models.

VGG16 correctly identified 370 images, with 53 false positives and 73 false negatives. ResNet-50 accurately predicted 289 out of 496 images but had 115 false negatives. InceptionV3 accurately predicted 360 images, excelling in identifying real images but with 80 false positives. Densenet-201 predicted 393 images correctly, showing strong performance with 33 false positives. The Ensemble Model had the highest accuracy, correctly predicting 403 images, offering the best balance.

Overall, the Ensemble Model outperformed the others, followed closely by Densenet-201. ResNet-50 had the weakest performance due to high false negatives, while VGG16 and InceptionV3 had balanced but slightly less optimal results compared to the Ensemble and Densenet-201 models.

D. *Deepfake Detection Result*

Fig. 5 illustrates the true label and predicted label for each image, providing insights into the model's performance on the dataset.

Fig. 5. Ensemble Models' prediction on the test set.

The ensemble model was assessed with a test set of 496 images. For an in-depth analysis, a subset of 9 images was selected to make predictions and visualize the outcomes. The model accurately predicted eight of these nine images, with only one image being misclassified.

VII. Conclusion and Future Work

Fraudsters have figured out how to outsmart traditional detection methods that usually depend on human judgment. Therefore, it's crucial to implement robust safeguards to detect and prevent deepfake-enabled fraud. In this study, we have developed a deepfake detection model using transfer learning and ensemble technique. By leveraging pre-trained models (ResNet50, VGG16, DenseNet201, InceptionV3) and combining their outputs, we achieved superior accuracy compared to individual models. In future, we wish to enhance the detection model by incorporating more sophisticated ensemble methods as well as extending the detection capabilities to include audio and video modalities.

Acknowledgment: The research work presented in this paper is an outcome of the M.Sc. thesis work of the first author which is funded by the NST fellowship under the Ministry of Science and Technology, Dhaka, Bangladesh in the fiscal year 2023-2024. We also thank the experts and personnel in relation to this research work.

References

[1] Atwan, J., Wedyan, M.O., Albashish, D., Aljaafrah, E., Alturki, R., & Alshawi, B. (2024). Using Deep Learning to Recognize Fake Faces. *International Journal of Advanced Computer Science and Applications*.

[2] Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, *53*(4), 3974-4026.

[3] Gambín, Á. F., Yazidi, A., Vasilakos, A., Haugerud, H., & Djenouri, Y. (2024). Deepfakes: current and future trends. *Artificial Intelligence Review*, *57*(3), 64.

[4] Fernando, T., Fookes, C., Denman, S., & Sridharan, S. (2019). Exploiting human social cognition for the detection of fake and fraudulent faces via memory networks. *arXiv preprint arXiv:1911.07844*.

[5] Guarnera, L., Giudice, O., & Battiato, S. (2020). Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 666-667).

[6] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5001-5010).

[7] Xu, B., Liu, J., Liang, J., Lu, W., & Zhang, Y. (2021). DeepFake Videos Detection Based on Texture Features. *Computers, Materials & Continua*, *68*(1).

[8] Zhang, X., Karaman, S., & Chang, S. F. (2019, December). Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-6). IEEE.

[9] Tariq, Shahroz & Lee, Sangyup & Kim, Hoyoung & Shin, Youjin & Woo, Simon. (2018). Detecting Both Machine and Human Created Fake Face Images In the Wild. 81-87. 10.1145/3267357.3267367.

[10] Ismail, A., Elpeltagy, M., S. Zaki, M., & Eldahshan, K. (2021). A new deep learning-based methodology for video deepfake detection using XGBoost. *Sensors*, *21*(16), 5413.

[11] Xuan, X., Peng, B., Wang, W., & Dong, J. (2019, October). On the generalization of GAN image forensics. In *Chinese conference on biometric recognition* (pp. 134-141). Springer International Publishing.

[12] Wang, S. Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8695-8704).

[13] Pan, S.J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*, 1345-1359.

[14] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[15] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[16] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).

[17] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

[18] Rokach, Lior. (2010). Ensemble-based classifiers. Artif. Intell. Rev.. 33. 1-39. 10.1007/s10462-009-9124-7.

[19] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401-4410).

[20] Xhlulu, "140k real and fake faces," 2020, accessed: 2024-06-19. [Online]. Available: https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces

[21] NVIDIA Research, "FFHQ Dataset," 2019, [Online]. Available: https://github.com/NVlabs/ffhq-dataset

[22] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee.

[23] Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

[24] Dubey, S. R., Singh, S. K., & Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, *503*, 92-108.

[25] Jadon, S. (2020, October). A survey of loss functions for semantic segmentation. In *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)* (pp. 1-7). IEEE.

[26] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Biography

Jannatul Mawa is currently going to complete her Master's Thesis in Computer Science and Engineering Department of Jahangirnagar University of the year 2022. She completed her Bachelor of Science (Hons.) degree in the same department from the same university.

Md. Humayun Kabir received his Doctor of Philosophy degree in the area of formal software development from the School of Computing, Dublin City University in 2007, Ireland, Master of Science degree in Computer Science from Dhaka University in 1992, Bachelor of Science (Hons.) degree in Applied Physics and Electronics from Dhaka University in 1991, Bangladesh. He is currently working as a Professor in the Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka. He worked as the Chairman of this department. His research interests include data mining, program transformation, program verification, program construction, software architecture and so on. He has been working for more than 20 years in private and public Universities as a teaching faculty. He wrote some Journal and conference papers, faculty research project reports and reviewed few papers for some Journals and conferences. He wrote some articles for computer magazine. He worked as the Editor-in-Chief for the Journal of the Department, and worked in the IWCI 2016 committee. His personal research interest include ICT-based development policy making for the betterment of the humanity.