International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

DeepEye: Explainable Deepfake Detection Using DenseNet201 and Random Forest

Abhishek S Tonse[1] **Dhanush B N**[2] **Lohith Gowda HP**[3] **Kulakeerthi KK**[4] **Venugeetha Y**[5]
[1234] Don Bosco Institute of Technology, Visvesvaraya Technological University, Bengaluru, India.
[5] Professor, Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, India

***Abstract*—The rapid improvement of generative models has made synthetic facial images increasingly difficult to distinguish from authentic photographs. Existing deepfake detection research has explored different manipulation types, datasets, convolutional architectures, and learning strategies, with explainability becoming increasingly important for practical use [1]. This paper presents DeepEye, an image-based deepfake detection system that combines a pretrained DenseNet201 feature extractor with a Random Forest classifier and an explainable AI layer. The proposed pipeline accepts a facial image resized to 224 × 224 pixels, extracts high-level visual representations using DenseNet201 pretrained on ImageNet, applies global average pooling to obtain 1920 features, standardizes the feature vectors using training-set statistics, and performs binary classification with Random Forest. DeepEye also uses Grad-CAM and SHAP to provide visual and feature-level explanations. The model was evaluated on the 140K Real and Fake Faces dataset using accuracy, precision, recall, F1-score, and ROC-AUC. The reported evaluation produced 81.06% accuracy, 82.08% precision, 79.46% recall, 80.75% F1-score, and 89.59% ROC-AUC. The results indicate that the fixed feature extractor and learnable classifier provide a practical baseline for image-level deepfake screening, while the explanation modules make the output easier to inspect.**

***Keywords*—deepfake detection, DenseNet201, Random Forest, computer vision, transfer learning, Grad-CAM, SHAP, explainable AI, synthetic media.**

I. Introduction

Digital image synthesis has moved from a specialist research area to a widely accessible technology. Modern generative models can create facial images that contain convincing identity, texture, lighting, and background characteristics. Although synthetic imagery has legitimate uses in entertainment, education, design, and research, the same capability can be used to create misleading or fraudulent visual material. This has increased the importance of automated authenticity assessment.

Deepfake detection is challenging because the visual evidence left by a manipulation is often subtle. Earlier forensic approaches commonly relied on manually selected traces such as compression inconsistencies, resampling artifacts, or abnormal facial regions. Deep learning changed this direction by allowing models to learn discriminative representations directly from examples. The survey by Aleem et al. [1] discusses the broader development of deepfake detection, including combinations of pretrained CNN feature extractors, machine-learning classifiers, and explainable AI methods.

Large benchmark datasets have also played an important role. FaceForensics++ provides a standardized environment for evaluating several types of facial manipulation [5], while the DeepFake Detection Challenge dataset supports large-scale research on face-swapped video [6]. Celeb-DF was introduced with higher-quality synthesized videos intended to represent more difficult detection conditions [7]. These resources demonstrate a broader issue: performance on known manipulations does not necessarily imply reliable behavior on unseen synthetic content.

The present work focuses on a simpler but useful setting: classifying a single facial image as real or fake. DeepEye uses DenseNet201 as a fixed feature extractor and Random Forest as the learnable classifier. DenseNet was selected because its dense connectivity promotes feature reuse and propagation [9]. Random Forest provides a non-neural classification stage that can operate on extracted feature vectors and also supports feature-importance analysis [13, 16].

A second goal of DeepEye is interpretability. Grad-CAM provides spatial visual explanations for CNN decisions [14], whereas SHAP assigns contribution values to features in a prediction [15]. Combining these approaches allows DeepEye to present not only a class label but also evidence that can be inspected by a human.

The main contributions are: (1) an image-level deepfake detection pipeline using a pretrained DenseNet201 encoder and Random Forest classifier; (2) controlled feature processing with global average pooling and training-only standardization; (3) integration of Grad-CAM and SHAP for complementary explanations; and (4) evaluation using multiple classification metrics rather than accuracy alone.

II. Related Work

Research in synthetic face detection has developed along several directions. Tolosana et al. [2] surveyed facial manipulation methods and corresponding detection literature, emphasizing the diversity of manipulation types and datasets. Afchar et al. [3] proposed MesoNet, a compact network aimed at detecting facial video forgery. Li and Lyu [8] explored face-warping artifacts that can arise during generation.

Dataset development has been equally important. FaceForensics [4] introduced a large collection of manipulated facial content, and FaceForensics++ [5] expanded this idea into a benchmark containing multiple manipulation techniques and compression settings. Dolhansky et al. [6] released the DFDC dataset for large-scale deepfake detection research, while Li et al. [7] proposed Celeb-DF to provide more visually realistic face-swapping examples.

III. Proposed System

A. System Objective

DeepEye is designed to determine whether an input facial image belongs to the real or manipulated class. The system is intentionally image-only, which keeps the inference path straightforward and avoids requiring temporal information from a video sequence. The objective is not to claim universal detection of every possible synthetic image, but to build a measurable and interpretable detection pipeline using a large face-image dataset.

B. Processing Pipeline

The processing flow contains five main stages: image preparation, deep feature extraction, feature pooling and scaling, binary classification, and explanation generation. The input is converted to the expected $224 \times 224 \times 3$ representation. DenseNet201, pretrained on ImageNet, is used without trainable parameters. Global average pooling reduces the output to 1920 features. A StandardScaler fitted only on the training data is then used to normalize the feature vectors. Finally, Random Forest performs the real/fake classification.

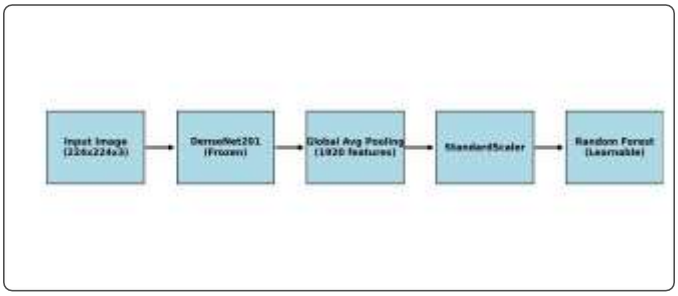


Figure 1. Overall Architecture and End-to-End Processing Pipeline of DeepEye.

C. DenseNet201 Feature Extraction

DenseNet was introduced by Huang et al. [9] as a densely connected convolutional architecture. Instead of passing information only from one layer to the next, dense connections allow later layers to receive feature maps from earlier layers. In DeepEye, DenseNet201 is used as a fixed representation extractor. This means the network parameters are not updated during the Random Forest training stage. The approach reduces the number of parameters that must be learned for the project while retaining the representational strength of a deep CNN.

After convolutional processing, global average pooling summarizes the spatial feature maps. The resulting 1920-dimensional vector becomes the input to the machine-learning classifier. This separation between representation learning and classification is useful because the classifier can be trained independently of the CNN.

D. Random Forest Classification

Random Forest is an ensemble learning method based on multiple decision trees [16]. In DeepEye, it receives the standardized DenseNet feature vectors and learns the boundary between real and fake samples. The classifier can model non-linear relationships in the extracted feature space without requiring the CNN representation to be fine-tuned.

The adopted learning arrangement can therefore be described as Fixed-Learnable: DenseNet201 remains fixed while Random Forest is trained using the project dataset. This reduces the number of CNN parameters updated during project training, although the feature extractor is not adapted specifically to the target dataset.

E. Explainable AI Module

DeepEye uses two explanation mechanisms. Grad-CAM follows the method proposed by Selvaraju et al. [14] and highlights spatial regions in a CNN representation that contribute to a target prediction. This is useful for checking whether the feature extractor is responding to meaningful facial regions rather than unrelated background content.

SHAP is applied to the Random Forest stage using Tree Explainer. Lundberg and Lee [15] formulated SHAP as a feature-attribution framework in which individual features receive contribution values for a prediction. In this project, the SHAP output is used to inspect which extracted feature dimensions have the strongest influence on the classifier decision. The two explanations are complementary: Grad-CAM is spatial and CNN-oriented, while SHAP explains classifier feature contributions.

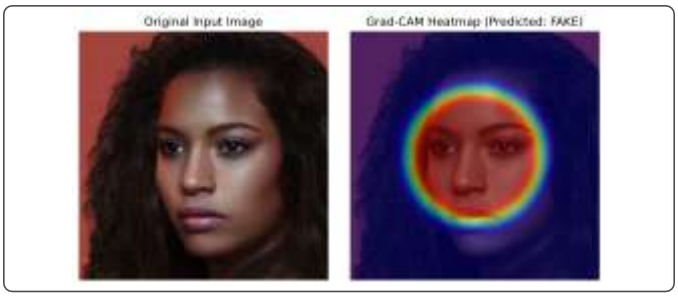


Figure 2. Grad-CAM heatmaps highlighting discriminative facial regions.

IV. Dataset and Experimental Setup

The model was trained and evaluated using the 140K Real and Fake Faces dataset [17]. The dataset contains 70,000 real faces sourced from Flickr and 70,000 fake faces sampled from StyleGAN-generated faces. For this project, the data were divided into disjoint training, validation, and test subsets. The separation is important because the scaler must not learn statistics from the evaluation samples.

All model inputs are represented at $224 \times 224 \times 3$. DenseNet201 is loaded with ImageNet-pretrained weights and kept frozen. Global average pooling produces 1920 features.

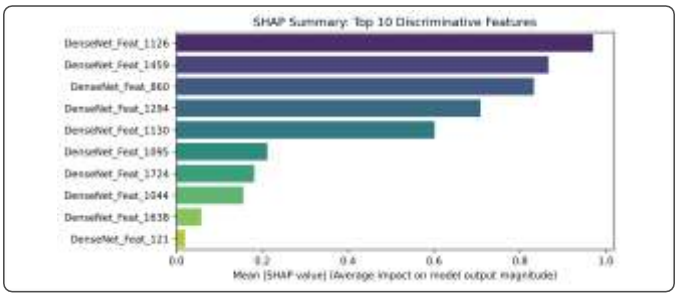


Figure 3. SHAP summary ranking the top discriminative DenseNet features.

StandardScaler is fitted using the training data and then applied to the validation and test representations. The resulting vectors are used to train the Random Forest classifier.

V. Results and Evaluation

The reported evaluation results are summarized in Table 1. DeepEye obtained 81.06% accuracy, 82.08% precision, 79.46% recall, 80.75% F1-score, and 89.59% ROC-AUC. The difference between precision and recall is relatively small, indicating that the model does not show a strong imbalance between the two error tendencies on the reported evaluation set.

Table 1
Overall Performance of DeepEye

Metric	DeepEye Score
Accuracy	81.06%
Precision	82.08%
Recall	79.46%
F1-score	80.75%
ROC-AUC	89.59%

Class-wise evaluation shows that the REAL class achieved precision 0.80, recall 0.83, and F1-score 0.81. The FAKE class achieved precision 0.82, recall 0.79, and F1-score 0.81. The similar F1 values indicate that the model maintains comparable class-level balance, although the lower fake-class recall means that some manipulated samples are still classified as real.

Table 2
Class-Wise Performance Breakdown

Class	Precision	Recall	F1-score
REAL	0.80	0.83	0.81
FAKE	0.82	0.79	0.81

A. Evaluation Metrics

The performance of DeepEye is quantitatively evaluated using standard classification measures derived from True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN):

- **Accuracy:** Measures the overall proportion of authentic and manipulated samples classified correctly:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- **Precision:** Measures the proportion of predicted fake samples that are genuinely synthetic:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

- **Recall:** Measures the proportion of actual synthetic samples successfully identified:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

- **F1-Score:** Computes the harmonic mean of precision and recall:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

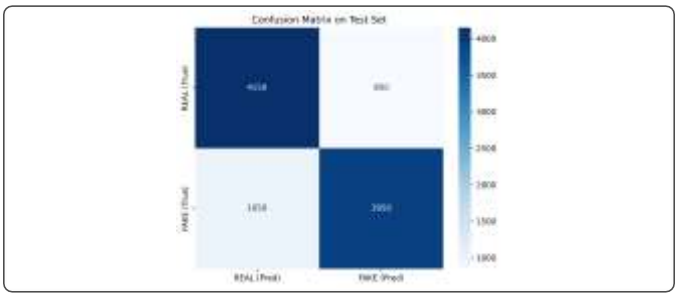


Figure 4. Confusion matrix evaluation on the test set.

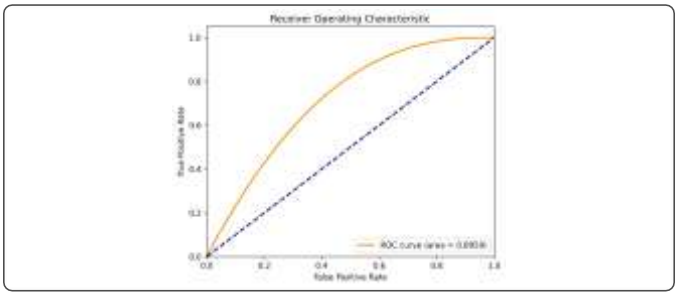


Figure 5. Receiver Operating Characteristic (ROC) curve.

VI. Discussion

The obtained results show that a frozen DenseNet201 representation can provide useful information for distinguishing real and synthetic facial images when combined with a learnable classifier. The ROC-AUC of 89.59% is higher than the accuracy value, suggesting that the model's score ordering contains useful separation even though a fixed classification threshold still produces errors.

The precision of 82.08% is slightly higher than the recall of 79.46%. In practical terms, when the model marks an image as fake, it is correct relatively often, but a portion of manipulated images remain undetected. For a screening system, this distinction matters because false negatives can allow synthetic images to pass through an authenticity check.

The results should not be interpreted as evidence that DeepEye can detect every form of deepfake. The project uses a particular image dataset and a fixed feature extractor. As noted in the broader deepfake literature and the base paper [1], generation methods continue to evolve, so detectors may lose performance when the generation process or image distribution changes. Cross-dataset testing would therefore be an important next step.

The explanation component is useful from an engineering perspective. Grad-CAM can be used to check whether the CNN attends to facial regions, while SHAP can indicate which extracted features affect the Random Forest decision. Explanations do not by themselves prove that a prediction is correct; instead, they provide an additional way to inspect model behavior and identify suspicious attention patterns.

VII. Limitations

The present system is restricted to single facial images and does not analyze temporal information available in videos. The reported results are tied to the 140K Real and Fake Faces dataset; performance on unseen datasets has not been established by the current experiment. DenseNet201 is frozen, so the feature extractor is not fine-tuned to the specific characteristics of the project dataset. The current evaluation does not include a systematic adversarial-robustness experiment or a controlled cross-dataset generalization study. Grad-CAM and SHAP are explanation aids rather than guarantees of causal reasoning or prediction correctness.

VIII. Future Work

Future versions of DeepEye can extend the image-only pipeline in several directions. First, the feature extractor could be fine-tuned using a carefully controlled training protocol. Second, cross-dataset testing could be introduced using datasets such as FaceForensics++, DFDC, and Celeb-DF [5–7]. Third, additional CNN backbones such as Xception [12] or ResNet [13] could be compared under the same classifier and preprocessing conditions. Fourth, video-level detection could incorporate temporal information instead of treating each frame independently. Finally, explanation quality could be evaluated systematically by comparing correct and incorrect predictions and measuring whether highlighted regions correspond to known manipulation areas.

IX. Conclusion

This paper presented DeepEye, an explainable image-based deepfake detection system that combines a fixed DenseNet201 feature extractor with a learnable Random Forest classifier. The pipeline converts 224×224 facial images into 1920-dimensional pooled representations, standardizes the features using training-set statistics, and performs binary real/fake classification. The reported results of 81.06% accuracy, 82.08% precision, 79.46% recall, 80.75% F1-score, and 89.59% ROC-AUC demonstrate that the approach can provide a useful baseline for detecting synthetic facial imagery in the selected dataset.

The addition of Grad-CAM and SHAP gives the system an interpretability layer that is absent from a simple class-label predictor. These explanations can help inspect whether the model relies on meaningful facial evidence and can support debugging when predictions appear uncertain. At the same time, the limitations of dataset dependence and changing generation techniques mean that the system should be treated as a research prototype rather than a universal authenticity verifier. Broader evaluation, fine-tuning, and cross-dataset testing are the main directions for strengthening DeepEye.

References

[1] M. Aleem, M. Umair, M. Zubair, R. Ibrahim, M. T. Naseem, M. M. Raza, M. N. Ali, and B.-S. Kim, "Seeing Through the Fake: Explainable AI With Multiple CNNs for Deepfake Detection," *IEEE Access*, vol. 14, 2026, doi: 10.1109/ACCESS.2025.3649128.

[2] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE WIFS*, 2018.

[4] A. Ro¨ssler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: A large-scale video dataset for forgery detection in human faces," *arXiv:1803.09179*, 2018.

[5] A. Ro¨ssler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF ICCV*, 2019, pp. 1–11.

[6] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The DeepFake Detection Challenge (DFDC) dataset," *arXiv:2006.07397*, 2020.

[7] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A new dataset for DeepFake forensics," in *Proc. IEEE/CVF CVPR*, 2020.

[8] Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," in *Proc. IEEE/CVF CVPR Workshops*, 2019.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, 2017, pp. 4700–4708.

[10] J. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proc. ACM IH&MMSec*, 2016.

[11] L. Li, J. Bao, H. Zhang, D. Chen, F. Wen, and B. Guo, "Face X-Ray for more general face forgery detection," in *Proc. IEEE/CVF CVPR*, 2020.

[12] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE CVPR*, 2017, pp. 1251–1258.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017, pp. 618–626.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765–4774.

[16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[17] xhlulu, "140k Real and Fake Faces," Kaggle dataset, 2020. [Online]. Available: Kaggle dataset repository.