A Hybrid Deep Learning Framework with Weighted Voting Ensemble of DenseNet and ResNet for watermark Text Classification

Bharathi Pilar[1], Safnaz[2]

[1,2]*Department of Computer Science, University College Mangalore*

Abstract- In recent years, the proliferation of digital content has led to increased concerns about intellectual property protection and content authenticity. Digital watermarks, especially those embedded as text in scene images, have emerged as a vital tool for asserting ownership and preventing unauthorized use. However, detecting and classifying such watermarked text in complex natural scenes remains a challenging task due to varying backgrounds, fonts, distortions and noise.

This research focuses on the classification of watermarked and non-watermarked text in scene images. To achieve this, we used deep convolutional neural networks, DenseNet-169, Squeeze-and- Excitation (SE) -ResNet and Wide ResNet, for robust feature extraction. The extracted features were then passed to Machine Learning (ML) classifiers, including Random Forest (RF), Support Vector Ma- chine (SVM), and Logistic Regression (LR), to per- form the final classification. In addition, a weighted voting mechanism was implemented to combine the outputs of individual classifiers and improve over- all accuracy. The proposed approach demonstrates promising results in distinguishing between water- marked and non-watermarked text, contributing to the advancement of automated watermark detection systems.

The proposed method shows promising results, with the highest accuracy of 97.13% achieved using SE- ResNet with weighted voting, followed by 95.08% using DenseNet201. This highlights SE-ResNet's effectiveness in ensemble classification for watermark detection.

Index Terms—deep learning, watermark text classification, natural scene images, weighted voting mechanism

1 INTRODUCTION

Digital images play a crucial role in conveying information in today's society. However, with technological advancements, images can be altered in ways that are nearly undetectable, which can be misused for harmful purposes such as spreading fake news and online rumors. Therefore, identifying manipulated areas in an image is vital to ensuring its authenticity [1].

Images are a key medium for information exchange in domains such as e-commerce and social networks, where they are widely used and rapidly shared. In modern digital life, many online images include visible watermarks to indicate ownership. To prevent unauthorized use of copyrighted content, it is essential to detect watermarks before further use. Consequently, developing an automated and accurate watermark detector has become increasingly important. Moreover, since visible watermarks play an important role in copyright protection, researchers have also explored methods for watermark removal after detection to evaluate their resilience and effectiveness [2].

Developing robust methods for visible watermark detection and removal remains a challenging task due to the diverse nature of visible watermarks as shown in Fig. 1. These watermarks may consist of text, symbols, or graphics, making it difficult to extract discriminative features from such varied patterns. In addition, variations in the shape, location, transparency, and size of the watermark on different types of watermarked images further complicate the task of accurately estimating watermark regions in real-world scenarios [3].

Scene text understanding has become a prominent area of research in computer vision due to its practical applications in automated navigation, content filtering, and multimedia authentication [4]. Recognizing watermarked text within natural scenes introduces additional complexity, as the text may be embedded with low contrast or obscured by background elements.

Traditional image processing techniques for watermark detection often depend on hand-crafted features and threshold-based approaches, which perform

inadequately under complex, real-world conditions. In contrast, recent advances in deep learning have significantly improved feature extraction and classification, especially when working with noisy or cluttered backgrounds [6].

Figure 1: Sample images for Watermarked and non watermarked text [19].

To handle the variability in watermark appearances, deep feature extractors such as DenseNet and ResNet variants provide more robust representation learning. Furthermore, ensemble learning methods, particularly those that use weighted voting mechanisms, have demonstrated notable success in improving classification accuracy in heterogeneous data sets [10]. Despite existing efforts in watermark detection, limited research has addressed the classification of watermarked versus non-watermarked text in natural scene images using ensemble-based deep learning approaches. This study aims to fill that gap by proposing a hybrid framework that integrates state-of-the-art convolutional neural networks with traditional classifiers, enhanced by a weighted voting mechanism, to achieve highly accurate classification performance.

The rest of this paper is organized as follows: Section 2 provides a review of existing literature related to watermark text classification, normal text classification. Section 3 outlines the proposed methodology, including pre-processing, feature extractions, ML classifiers, the weighted voting mechanism, and the proposed model architectures. Section 4 discusses the experimental results with dataset description, evaluation metrics. Finally, Section 5 concludes the paper by summarizing the main insights and highlighting the implications of this study.

2 RELATED WORK AND BACKGROUND OF THE STUDY

In recent years, watermark detection and classification in natural scene images have gained considerable attention due to the widespread use of digital content and the growing need to protect intellectual property. Various deep learning models and traditional ML techniques have been explored to tackle this challenge effectively.

Convolutional Neural Networks (CNNs) have emerged as powerful tools for feature extraction in image-based classification tasks. Architectures such as InceptionV3, DenseNet, and ResNet have demonstrated excellent performance in capturing detailed and hierarchical image features, making them suitable for distinguishing between watermarked and non-watermarked content. Researchers have also experimented with hybrid models, where deep features are extracted using CNNs and subsequently classified using ML algorithms like SVM, k-NN, and RF, achieving notable improvements in accuracy.

Furthermore, ensemble learning techniques, particularly weighted voting mechanisms, have been introduced to combine the strengths of multiple classifiers.

These methods enhance the reliability and robustness of predictions by leveraging the individual performance metrics of each model. Collectively, these prior works lay the groundwork for developing more accurate and generalized watermark detection systems. Bharathi Pilar and Safnaz, [9] developed a model for watermarked and non-watermarked text classification using InceptionV3 model with ML classifiers k-NN, SVM and RF. Also a weighted voting decision mechanism was employed, leveraging the strengths of each classifier for enhancing the classification performance. To overcome the limitations of classifying entire documents, which is ineffective for mixed-source texts containing only partial watermarks, Y. Xu, et al., [7] developed the Geometric Cover Detector (GCD). This classification framework checks for watermark signals across various text segments defined by a geometric cover. A positive classification for the whole document is triggered if even one segment is identified as watermarked, providing a more robust approach than baseline methods for detecting the presence of LLM-generated text in longer documents.

In their word-shift watermarking approach, Young-Won Kim, [5] introduced classification as a method to organize text for embedding. They classify individual words using width features and subsequently classify text segments based on these word classes. This allows them to group segments sharing the same class label, distributed throughout the document. Watermark bits are then encoded by manipulating inter-word space statistics collectively for all segments within a given class, rather than modifying individual words or lines independently based on the watermark bit directly.

Many researchers have underscored the importance of watermarks in safeguarding data integrity, verifying content authenticity, identifying ownership, and protecting copyrights. T. Van Phan and M. Nakagawa, [18] proposed a distinctive method for online classification of handwritten documents into text and non-text categories by employing Recurrent Neural Networks (RNN) and an improved Long Short-Term Memory (LSTM) network. Their model classifies strokes in digital ink documents by capturing both local and global contextual cues, thereby enhancing classification accuracy through an efficient yet simple strategy that leverages temporal relationships among strokes. Safnaz and et al. [11] presented an Xcpetion-based approach for classification of watermarked and non-watermarked text in natural scene images. The performance is evaluated using a comprehensive range of metrics, including accuracy, precision, recall, and F1 score, resulting in respective values of 86.5%, 84.76%, 89%, and 86.82%.

Addressing the challenge of distinguishing text from non-text in freehand online documents, A. Delaye and C.-L. Liu, [16] introduced a Conditional Random Field (CRF)-based approach. By jointly modeling multiple information sources—including local, spatial, and temporal contexts—this method improves stroke-level classification precision.

N. Sharma and et al., [20] presented an innovative component-based technique for categorizing text frames. Their method starts by detecting potential word clusters based on gradient information derived from RGB video frames. Sobel edge features from these clusters are extracted for further analysis. The approach eliminates erroneous components before applying linearity checks, where the linear alignment of component centroids is examined. If the components meet the defined linearity condition, the frame is marked as text; otherwise, it is identified as non-text.

The robustness of the method across different text orientations is validated through tests on a diverse dataset.

Yuan Zhao and et al, [8] utilized Markov Random Fields (MRFs) to differentiate between text and non-text ink strokes in digitized handwritten Japanese documents. By effectively modeling spatial dependencies among strokes, and training SVM classifiers to assess both individual and paired strokes, they converted SVM outputs into probabilistic clique potentials for the MRF. Their evaluation on the TUAT Kon-date dataset revealed that this method outperformed models that relied solely on individual strokes or Hidden Markov Models (HMMs) for sequence classification.

A. Delaye and C.-L. Liu, [17] further explored CRF models to enrich classification accuracy by incorporating contextual knowledge during the classification of text and non-text strokes in online handwritten content. Their approach effectively combined spatial and temporal interactions for more comprehensive context modeling in the classification process.

T. Van Phan and M. Nakagawa, [18] developed a sophisticated classification model aimed at distinguishing text from non-text strokes in digital ink documents. Their architecture integrates both local and global con- texts, using a neural network framework that captures the overall structure and semantics of stroke sequences. Recent advancements in watermark detection and classification leverage deep learning, hybrid models, and ensemble techniques to enhance accuracy and robustness. These studies collectively provide a strong foundation for developing reliable systems to distinguish watermarked content in diverse contexts.

3 PROPOSED METHODOLOGY

The proposed framework for watermark and non-watermark text classification comprises several key stages: preprocessing, feature extraction using deep CNN architectures, classification using traditional ML models, and final decision making through a weighted voting mechanism. An overview of the methodology is illustrated in Fig. 2.
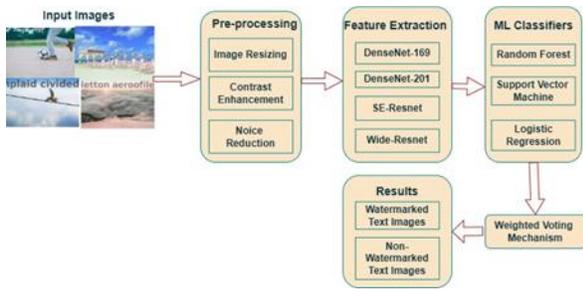
Figure 2: Architecture of proposed model.

3.1   Pre-processing

A) Image Resizing: All images were resized to a fixed dimension of 224×224 pixels, depending on the input requirements of the respective deep learning models used in the study.

B) Contrast Enhancement: Contrast enhancement techniques, such as histogram equalization and CLAHE (Contrast Limited Adaptive Histogram Equalization), were applied to improve the visibility of watermark text by enhancing the contrast between the text and the background.

C) Noise Reduction: Noise reduction was performed using filtering techniques such as Gaussian blur and median filtering to remove background artifacts and smooth the images, thereby improving the clarity of watermark text regions.

3.2   Feature Extraction

Feature extraction is a crucial step in the proposed framework, where deep convolutional neural networks are employed to automatically learn rich and discriminative features from watermarked and non-watermarked text images. This enables for accurate classification by capturing subtle variations in texture and structure. Here we have used DenseNet-169, 201 and SE-Resnet and Wide Resnet.

DenseNet-169

DenseNet-169, introduced by Huang et al. [12], is a convolutional neural network that connects each layer to every other layer in a feed-forward manner as shown in Fig. 3. This dense connectivity pattern promotes feature reuse, strengthens gradient flow, and allows for efficient learning with fewer parameters. In this work, DenseNet-169 was used as a feature extractor to effectively capture fine-grained visual features from water- marked and non-watermarked text images.

3.2.1   DenseNet-201

introduced by Huang et al. [12]. It is a deeper variant of DenseNet as shown in Fig. 3, follows the same dense connectivity principle but with an increased number of layers. This results in improved feature representation and learning of more complex image structures. It was employed in our study to extract high-dimensional features that are essential for distinguishing between different watermarking patterns in scene images.
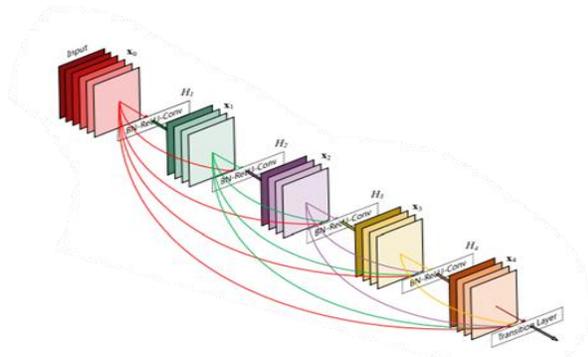

Figure 3: Architecture of the Dense Blocks [12]

3.2.2   SE-ResNet

SE-ResNet proposed by Hu et al. [13], enhances the ResNet architecture by incorporating SE blocks. These blocks adaptively recalibrate channel-wise feature responses, allowing the model to emphasize informative features and suppress irrelevant ones. In the context of watermark classification, SE-ResNet demonstrated strong performance in extracting robust features from variable watermark styles. The SE block integration is illustrated in Fig. 4.

3.2.3   Wide ResNet

Wide ResNet, introduced by Zagoruyko and Komodakis [14], modifies the traditional ResNet by increasing the width of residual blocks instead of depth. This approach reduces training time and prevents overfitting while retaining the benefits of residual learning. Wide ResNet
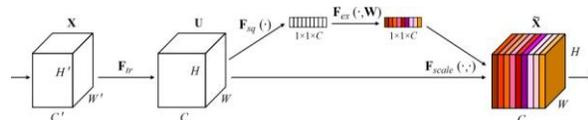

Figure 4: A Squeeze-and-Excitation block [13]

was utilized in this study for its ability to capture broad

features efficiently, making it suitable for datasets with complex variations. The model structure is illustrated in Fig. 5.
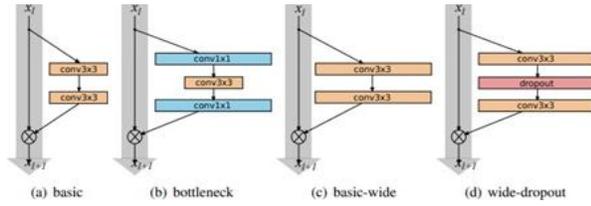

(a) basic   (b) bottleneck   (c) basic-wide   (d) wide-dropout

Figure 5: A Various residual blocks used in the paper. Batch normalization and ReLU precede each convolution [14]

3.3 ML Classifiers

The extracted deep features are then passed to three different ML classifiers:

• RF: It is an ensemble learning method based on decision trees. It improves classification performance by building multiple decision trees during training and aggregating their outputs. In the con- text of this research, RF was employed to process deep features extracted from CNN backbones. Its ability to handle high-dimensional data and reduce overfitting through averaging makes it a powerful classifier for complex watermark detection tasks. RF also contributes significantly in the weighted voting ensemble by providing robust predictions even under data variability.

• SVM: It is a supervised learning algorithm that aims to find the optimal hyperplane that maximally separates the classes in the feature space. In this work, SVM was used to classify the extracted CNN features of scene text images. It is particularly effective in high-dimensional spaces and per- forms well when there's a clear margin of separation between classes. Given the subtle visual cues between watermarked and non-watermarked text, SVM helps enhance classification precision and forms a critical part of the ensemble decision mechanism.

  • LR: It is a linear classifier widely used for bi-nary classification tasks. In this study, LR serves as a baseline ML model for distinguishing be- tween watermarked and non-watermarked text. It operates by mapping the high-dimensional features extracted from deep neural networks (such as DenseNet169, DenseNet201, SE-ResNet,

and Wide ResNet) into a probabilistic output that estimates the likelihood of an image belonging to either class. Despite its simplicity, LR is effective when the features are well-separated and of- ten complements more complex models within an ensemble framework.

3.4 Weighted voting mechanism

In ensemble learning, the weighted voting mechanism is an effective method for combining predictions from multiple classifiers to improve overall performance. Unlike simple majority voting, where each classifier has equal influence, weighted voting assigns importance (weights) to each classifier based on its performance metrics such as accuracy, precision, or F1-score on validation data.

In our research, we used a weighted voting ensemble combining the outputs of LR, RF, SVM. Each classifier was trained on features extracted from deep CNN models (DenseNet-169, DenseNet-201, SE-ResNet, and Wide ResNet), and then the final prediction for each im- age was made based on a weighted sum of the individual classifier predictions. This approach leverages the strengths of each model, ensuring more balanced and accurate classification between watermarked and non- watermarked text images.

The prediction score P for a class label is computed using the weighted sum of individual classifier outputs [15]:

$$P(y = c) = \sum_{i=1}^{n} w_i \cdot h_i(x)$$

Where,

• $P(y = c)$ is the weighted prediction score for class c,
• $h_i(x)$ is the prediction of the ith classifier for input x,
• w_i is the weight assigned to the ith classifier (e.g., based on F1-score or validation accuracy),
• n is the total number of classifiers.

The final predicted class is then determined by:

$$\hat{y} = \arg \max_c P(y = c)$$

This method enhances robustness by reducing the influence of weaker models and emphasizing those with stronger generalization capabilities, leading to improved accuracy across both datasets used in our experiments.

4 RESULTS AND DISCUSSION

The experiments were carried out on a system equipped with an Intel Core i5-1135G7 processor (2.40 GHz), 8 GB RAM, and a 4 GB GPU, running a 64-bit version of Windows 11.

4.1 Dataset Description

The dataset was sourced from publicly available Kaggle repositories—Watermarked / Not Watermarked Images [19] which were specifically created for watermark detection and classification tasks. The original dataset contained 15,000 images. For the purpose of analysis, a balanced subset of 3,000 images from each category—watermarked and non-watermarked—was selected. The dataset was then divided into 80% for training and 20% for testing.

4.2 Evaluation Metrics

To assess the model's performance, metrics such as accuracy 1, precision 2, recall 3, and F1-score 4 were utilized. The primary objective of this research is to enhance accuracy while preserving image information during the training process. Recall evaluates the model's capability to correctly identify actual positive instances, while precision reflects the proportion of true positive predictions among all samples labeled as positive. Accuracy indicates the overall reliability of the model by considering both correctly predicted positive and negative cases. The F1-score, which combines precision and recall, provides a balanced measure of the model's ability to differentiate between watermarked

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

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

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

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

4.3 Results

The classification performance of various deep feature extractors combined with traditional classifiers is presented in Tables 1 and 2. Table1 illustrates the results using DenseNet169 and DenseNet201 as feature extractors, while Table 2 shows the performance of SE-ResNet and Wide-ResNet. Models based on SE-ResNet and Wide-ResNet consistently outperform the DenseNet-based models across all evaluation metrics—accuracy, precision, recall, and F1-score.

The best-performing configuration is SE-ResNet with LR, achieving an accuracy of 96.20% and an F1-score of 96.00%, surpassing the top DenseNet201 + LR setup, which recorded 90.56% accuracy and 90.99% F1-score. The ensemble weighted voting results further confirm this trend, with SE-ResNet and Wide-ResNet ensembles yielding the highest overall performance. SE-ResNet achieved 97.13% accuracy and an F1-score of 97.09%, while Wide-ResNet reached 94.62% accuracy and an F1-score of 93.13% as shown in Table 2.

This comparative analysis demonstrates that residual-based architectures, particularly SE-ResNet, offer more effective and discriminative feature representations for watermark classification tasks compared to DenseNet variants.

Table 1: Performance Comparison Using DenseNet169 and DenseNet201 Feature Extractors

DATASET	FEATURE EXTRACTOR	CLASSIFIER	ACCURACY	PRECISION	RECALL	F1 SCORE
Watermarked / Not Watermarked Images	DenseNet169	RF	87.86	87.91	87.96	87.35
		SVM	88.08	88.15	88.12	88.49
		LR	89.36	89.35	89.96	89.75
		WEIGHTED	93.43	93.47	93.68	93.58
	DenseNet201	RF	89.33	89.34	89.34	89.33
		SVM	90.34	89.02	88.51	89.66
		LR	90.56	91.43	90.18	90.99
		WEIGHTED	95.08	94.93	94.34	94.99

Table 2: Performance Comparison Using SE-ResNet and Wide-ResNet Feature Extractors

DATASET	FEATURE EXTRACTOR	CLASSIFIER	ACCURACY	PRECISION	RECALL	F1 SCORE
Watermarked / Not Watermarked Images	SE-ResNet	RF	94.12	94.26	94.17	94.21
		SVM	95.08	95.07	95.08	95.05
		LR	96.20	96.66	96.98	96.00
		WEIGHTED	97.13	97.33	97.41	97.09
	Wide-ResNet	RF	91.64	90.70	91.88	91.06
		SVM	92.45	91.99	91.77	90.27
		LR	93.76	92.89	93.93	92.05
		WEIGHTED	94.62	93.86	94.53	93.13

As part of our previous research, we have al- ready performed watermarked text classification using ResNet101V2 and DenseNet201 for the same dataset, and the results are presented in Table 3.

Table 3: Performance Comparison of Feature Extractors

Feature Extractor	Accuracy	Precision	Recall	F1 Score
ResNet101V2	86.24	85.88	88.31	87.08
DenseNet201	92.44	89.87	93.56	94.54

5 CONCLUSION

This study presents a deep ensemble learning framework for the classification of watermarked and non-watermarked text in scene images. The proposed approach leverages the strengths of multiple deep convolutional neural networks—DenseNet-169, DenseNet 201, SE-ResNet, and Wide ResNet—for feature extraction, which are then fed into classical machine learning classifiers, including RF, SVM, and LR. To enhance the final decision-making process, a weighted voting mechanism was employed, effectively improving classification performance by integrating the strengths of individual classifiers.

Experimental results demonstrate that the ensemble model achieves the highest accuracy of 97.13% using SE-ResNet, followed by 95.08% with DenseNet201, outperforming individual classifier setups. These results highlight the superior generalization capabilities of the ensemble model across diverse watermark patterns, backgrounds, and distortions. The proposed method also shows significant improvement over standalone models, proving the effectiveness of combining deep learning and ensemble strategies for watermark detection tasks.

The findings of this research highlight the potential of hybrid frameworks in solving complex classification problems related to image authenticity and copyright protection. Future work may focus on expanding the dataset, incorporating multilingual or stylized watermark text, and exploring real-time watermark detection and removal systems for broader deployment in digital content verification scenarios.

REFERENCE

[1] Yuan Zhao, Bo Liu, Tianqing Zhu, Ming Ding, Xin Yu, and Wanlei Zhou, "Proactive image manipulation detection via deep semi-fragile watermark," *Neurocomputing*, vol. 585, p. 127593, 2024.

[2] Danni Cheng, Xiang Li, Wei-Hong Li, Chan Lu, Fake Li, Hua Zhao, and Wei-Shi Zheng, "Large-scale visible watermark detection and removal with deep convolutional networks," in *Proc. Pattern Recognition and Computer Vision (PRCV)*, Guangzhou, China, Nov. 23–26, 2018, pp. 27–40.

[3] "Comprehensive review of watermark detection and recognition techniques: Challenges, applications and future perspectives," in *Futuristic Trends in Artificial Intelligence*, vol. 3, book 1, IIP Series, 2024.

[4] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Young-Won Kim, Kyung-Ae Moon, and Il-Seok Oh, "A text watermarking algorithm based on word classification and inter-word space statistics," in *Proc. Int. Conf. on Document Analysis and Recognition (ICDAR)*, pp. 775–779, 2003.

[6] Y. Zhao, et al., "Deep learning techniques for image forensics," *J. Vis. Commun. Image Represent.*, vol. 58, pp. 1–9, 2019.

[7] Xuandong Zhao, Chenwen Liao, Yu-Xiang Wang, and Lei Li, "Efficiently identifying watermarked segments in mixed-source texts," *arXiv preprint arXiv:2410.03600*, 2024.

[8] X.-D. Zhou and C.-L. Liu, "Text/non-text ink stroke classification in Japanese handwriting based on Markov random fields," in *Proc. Int. Conf. on Document Analysis and Recognition (ICDAR)*, vol. 1, IEEE, 2007, pp. 377–381.

[9] Bharathi Pilar, et al., "InceptionV3-driven multi-classifier voting system for watermark classification," in *Proc. 2025 Int. Conf. on Computational, Communication and Information Technology (IC- CCIT)*, pp. 471–476, IEEE, 2025.

[10] Y. Xu, et al., "Learning ensemble classifiers with weighted voting for image classification," *Pattern Recognition*, vol. 102, 2020.

[11] Safnaz, Bharathi Pilar, and Shivakumara Palaiahnakote, "Scene images: An Xception-based approach," in *Proc. Fifth Int. Conf. on Computing and Network Communications (CoCoNet 2023)*, vol. 1221, Springer Nature, 2025, p. 455.

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017.

[13] Jie Hu, Li Shen, and Gang Sun, "Squeeze- and-excitation networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018.

[14] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[15] C. Xu and Z. Wang, "Learning ensemble classifiers with weighted voting for robust classification," *IEEE Access*, vol. 8, pp. 43724–43735, 2020.

[16] A. Delaye and C.-L. Liu, "Text/non-text classification in online handwritten documents with con- ditional random fields," in *Proc. Chinese Conf. on Pattern Recognition (CCPR)*, vol. 5, Springer, 2012, pp. 514–521.

[17] A. Delaye and C.-L. Liu, "Contextual text/non-text stroke classification in online handwritten notes with conditional random fields," *Pattern Recognition*, vol. 47, no. 3, pp. 959–968, 2014.

[18] T. Van Phan and M. Nakagawa, "Combination of global and local contexts for text/non-text classification in heterogeneous online handwritten documents," *Pattern Recognition*, vol. 51, pp. 112–124, 2016.

[19] F.Pollano, "Watermarked/Not Water- marked Images," *Kaggle*, 2024. [Online]. Available: https://www.kaggle. com/datasets/felicepollano/ watermarked-not-watermarked-images

[20] N. Sharma, P. Shivakumara, U. Pal, M. Blumen-stein, and C. L. Tan, "Piece-wise linearity based method for text frame classification in video," *Pattern Recognition*, vol. 48, no. 3, pp. 862–881, 2015.