ISYE 6740 Final Report – Project Group 003

Team Member Names: (1) Lam Alan C, (2) Hung Sean and (3) Flynt Trinh N

Project Title: ***Detecting Document Forgery Using Machine Learning on Receipt Images***

1. Introduction and Problem Statement

Receipt forgery presents a growing threat across sectors such as finance, insurance, and government, where receipts serve as critical evidence for financial reporting, tax auditing, and claims processing. With the proliferation of advanced image editing and OCR (Optical Character Recognition) manipulation tools, it has become increasingly easy to fabricate or alter receipts, posing significant challenges for traditional, rule-based fraud detection systems.

Manual inspection of receipts is time-consuming, inconsistent, and susceptible to human error, particularly given the variability in receipt layouts, fonts, and printing styles. As a scalable and objective alternative, this project explores the use of machine learning to automate forgery detection in scanned receipts.

The primary objective of this project is to design and implement an interpretable, multimodal machine learning framework that accurately classifies financial receipts as authentic or forged. The system aims to leverage both visual and textual information to improve detection accuracy, address class imbalance, and ensure transparency in model decision-making.

The task is formulated as a supervised binary classification problem, where each receipt is labeled as either authentic or forged. The input data consists of scanned receipt images and their corresponding OCR-transcribed text, which was pre-provided in the dataset. As such, performing OCR is not part of the project scope, the focus lies in analyzing and modeling the provided text and image data to detect forgery.

To achieve this, multimodal features were extracted from both data types: semantic and statistical patterns from the visual domain, and structural and contextual cues from the textual domain. Given the interpretability needs in high-stakes domains like finance, the solution prioritizes transparency and explainability. Rather than relying solely on deep learning, the proposed approach integrates:

1. Hand-engineered visual and textual features
2. Pretrained models for contextual and residual analysis
3. Traditional classifiers and lightweight ensemble strategies

This hybrid framework balances performance and interpretability, ensuring that each prediction of forgery can be supported by meaningful, understandable evidence. The final model leverages late fusion of probabilistic outputs from image and text classifiers to improve fraud detection performance while maintaining trust and accountability in decision-making.

2. Data Source

2.1. Data Description and EDA

This project utilizes the *Find It Again!* dataset, introduced by Tornés et al. (2023), which is derived from the SROIE (Scanned Receipts OCR and Information Extraction) dataset and was also featured in the ICPR 2018 "Find It!" Fraud Detection Contest. The dataset provides a comprehensive foundation for developing both content-based and image-based receipt forgery detection models.

The dataset consists of 988 scanned receipt samples, each accompanied by:

- A high-resolution image of the receipt
- OCR-transcribed textual content
- Ground truth labels indicating authenticity (genuine or forged)
- Detailed annotations identifying fraudulent modifications, including bounding boxes and the specific types of altered entities (e.g., price, date, merchant name)

Of the 988 receipts, 163 have been realistically forged, either manually or through digital manipulation, while the remaining 825 are authentic. This dataset consists of both visual and textual data along with detailed forgery annotations, which makes it especially well-suited for building and evaluating machine learning models aimed at detecting subtle and diverse forms of document forgery.

The dataset is publicly available and can be accessed via the following link: Find It Again! Dataset on Kaggle [https://www.kaggle.com/datasets/nikita2998/find-it-again-dataset]

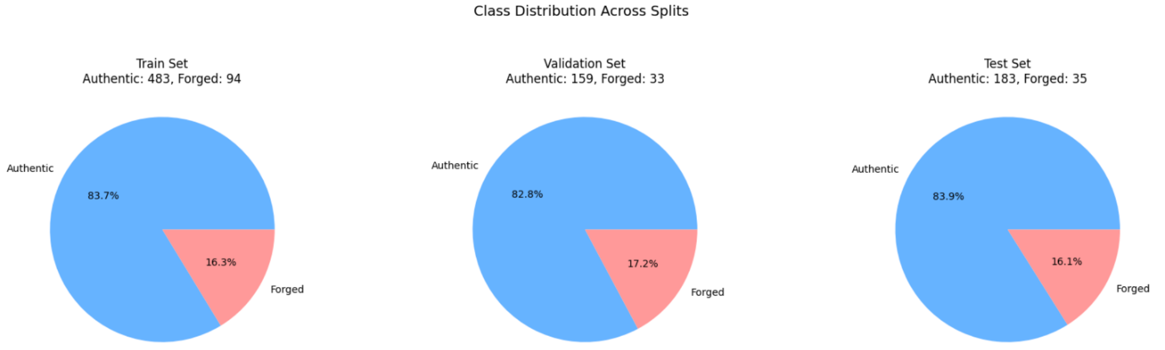


Figure 1: Basic Exploratory Data Analysis (EDA) on Train, Validation and Test dataset

With respect to Figure 1, it indicates a significant class imbalance in the dataset, which is likely to hinder the performance and reliability of machine learning models across training, validation, and testing phases. This observation aligns with feedbacks received on the project proposal, including suggestions such as: '*Only suggestion would be to briefly explain how you'll deal with the small number of forged samples'* and '*One thing that stands out to me as potentially problematic is the sample size of the dataset, it feels like having only 163 forgeries could be a limiting factor that leads to overfitting, although I could see how it could be difficult to garner a larger sample size with quality data.'*. This significant class imbalance in the dataset issue will be addressed in the methodology section.

Furthermore, after performing EDA and verifying the uniqueness of each dataset, one data point was identified as being present in both the training and validation sets. Specifically, the image and OCR files resided in the training folder, while the corresponding annotation appeared in both the training and validation annotations. To resolve this inconsistency, the data point was retained in the training set and removed from the validation set, resulting in a final total of 987 unique data points.

It is also noteworthy that there are a few different types of forgery classifications which consist of the following:

1. CPI = copy and paste inside the document
2. CUT = deletion of one or more characters/words
3. PIX = Pixel-level redraw / clone brush
4. IMI = creation of a text box imitating the font
5. CPO = copy and paste from another document
6. Autre = anything else. Flags may co-exist on one box.

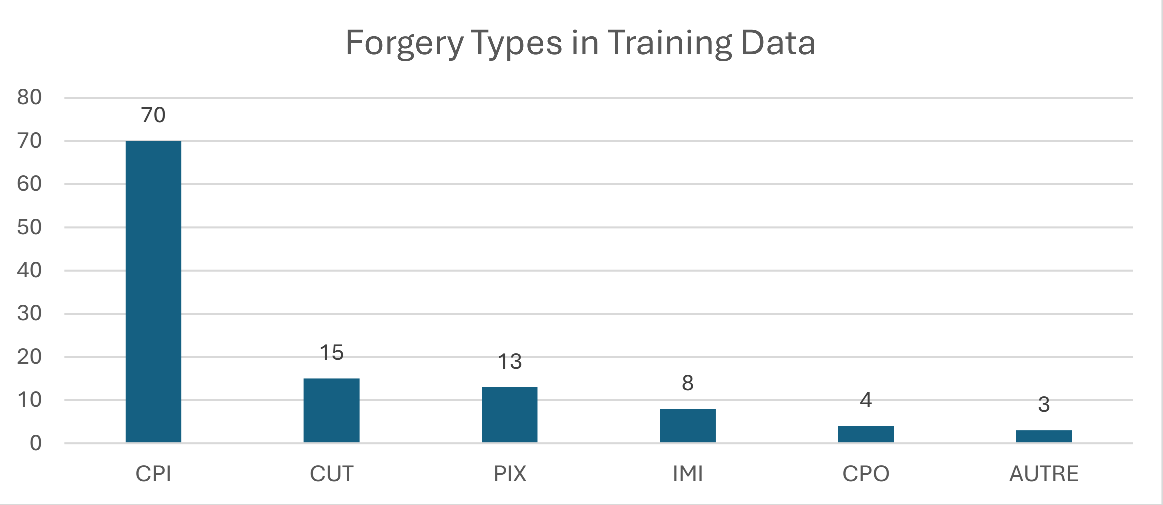


Figure 2: Forgery Types with respective Counts for Training Dataset

2.2. Data Challenges and Constraints

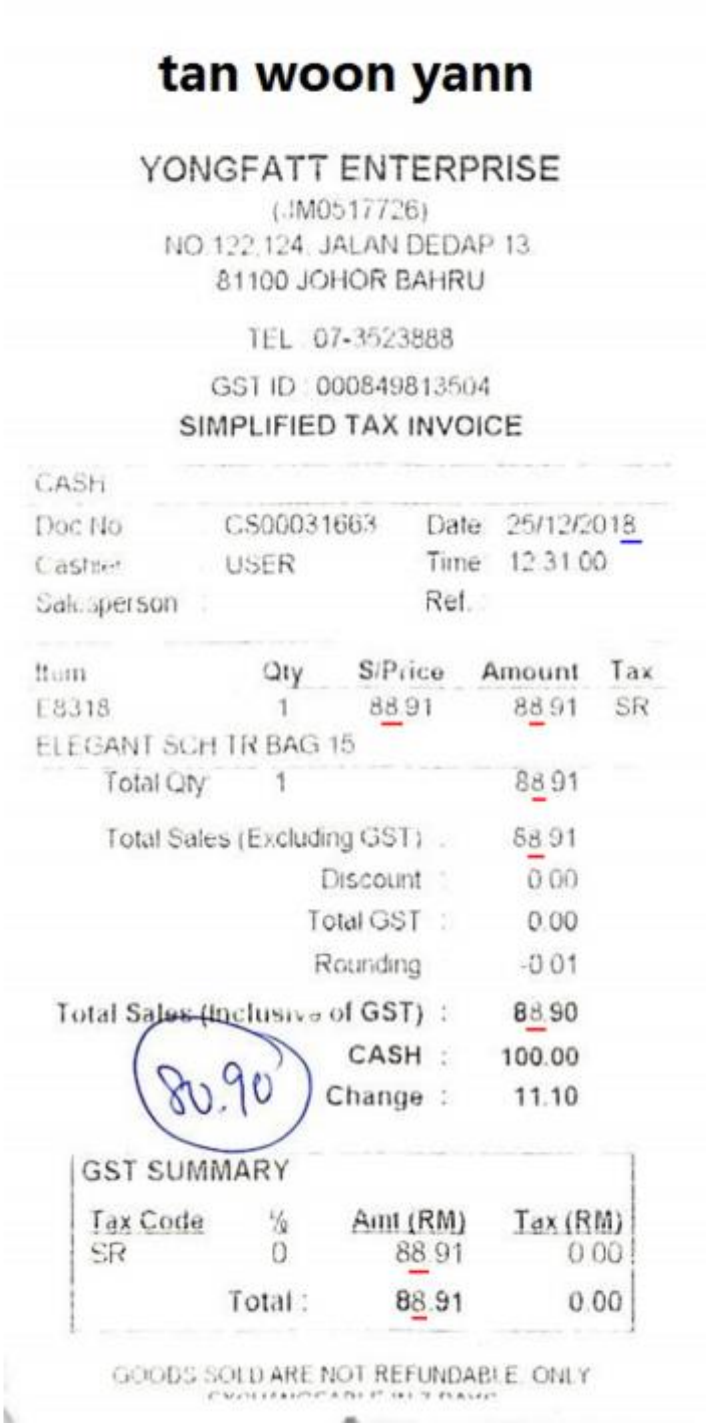
During the initial exploration of the dataset, several limitations and inconsistencies emerged that influenced both preprocessing pipeline and model strategy.

- Image Format Limitations: All image files were provided as PNG format, a lossless compression method. This posted a challenge for applying a common change detection for images, called Error Level Analysis (ELA), which relies on JPEG compression to reveal inconsistencies that may indicate tampering. As a result, there was an inability to pursue JPEG compression and ELA-based feature engineering.
- Visual Noise and Variability: The receipt images contained significant variability in format, lighting, and clarity. Some receipts included handwritten annotations, which introduced additional noise and potentially interfered with both translations and visual feature extraction. This variability required robust preprocessing and feature selection to minimize impact of irrelevant components.
- Dependence on Provided OCR Text: For the text data, a reliance exclusively on the OCR outputs provided with the dataset. While custom OCR pipelines could potentially improve translations quality, implementation would have been outside the scope of this project. Therefore, the provided text was as a fixed input and understanding that any translation errors would propagate through the entire text pipeline.
- Lack of Untampered Ground Truth for Forgeries: One of the more realistic, but still challenging aspects of the dataset was the absence of the unaltered versions for the tampered receipts. The design mirrors real-world fraud detection task, where the original copies may not be available. However, this aspect of the dataset still added a level of complexity to model training and evaluation, given that there was not a direct comparison between forged receipts against their real counterparts.

These considerations influenced several key decisions in the project, from choosing appropriate feature extraction techniques to managing noise and variation in the model evaluation.

1.1. Data Example

To further demonstrate a forged receipt, “X00016469622.png” is shown below and was identified as forged under the CPI category.



tan woon yann

YONGFATT ENTERPRISE
(JM0517726)
NO 122,124, JALAN DEDAP 13
81100 JOHOR BAHRU

TEL 07-3523888

GST ID : 000849813504
SIMPLIFIED TAX INVOICE

CASH

Doc No	CS00031663	Date	25/12/2018
Cashier	USER	Time	12:31:00
Salesperson :		Ref. :	

Item	Qty	S/Price	Amount	Tax
E8318	1	88.91	88.91	SR
ELEGANT SCH TR BAG 15				
Total Qty	1		88.91	

Total Sales (Excluding GST) :	88.91
Discount :	0.00
Total GST :	0.00
Rounding	-0.01
Total Sales (Inclusive of GST) :	**88.90**
CASH :	**100.00**
Change :	**11.10**

80.90

GST SUMMARY

Tax Code	%	Amt (RM)	Tax (RM)
SR	0	88.91	0.00
Total :		88.91	0.00

GOODS SOLD ARE NOT REFUNDABLE, ONLY

Figure 3: Forged receipt example ‘X00016469622.png’

3. Methodology

From a high-level perspective, Figure 4 depicts the respective types of data file utilized and steps to achieve the final creation of the hybrid model.

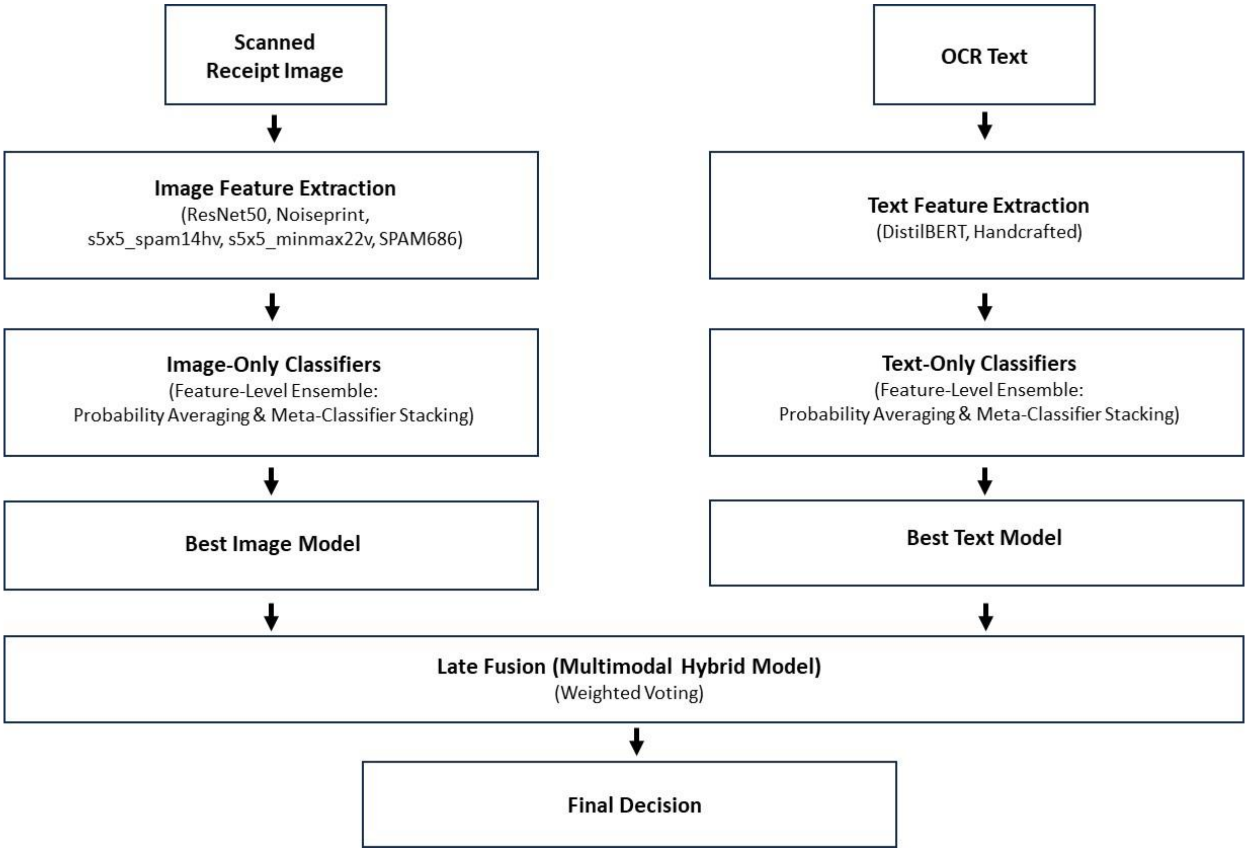

Scanned Receipt Image
OCR Text
Image Feature Extraction
(ResNet50, Noiseprint, s5x5_spam14hv, s5x5_minmax22v, SPAM686)
Text Feature Extraction
(DistilBERT, Handcrafted)
Image-Only Classifiers
(Feature-Level Ensemble: Probability Averaging & Meta-Classifier Stacking)
Text-Only Classifiers
(Feature-Level Ensemble: Probability Averaging & Meta-Classifier Stacking)
Best Image Model
Best Text Model
Late Fusion (Multimodal Hybrid Model)
(Weighted Voting)
Final Decision

Figure 4: Process Flow Diagram for Single file type modelling to Late Fusion modelling

3.1. Feature Engineering

Feature engineering was approached by independently extracting characteristics from the image and text files. This strategy enabled the development of separate models tailored to each file type, which were later integrated into a hybrid model. To provide a clearer understanding of the extraction methods explored, the two file types, image and text, are discussed separately, along with the specific techniques applied to each.

3.2. Image Feature Extraction

To capture visual cues relevant to forgery detection, three complementary approaches were used to extract features from receipt images: (1) deep neural embeddings, (2) sensor noise residuals, and (3) hand-engineered statistical features. These methods were organized into three main categories, as follows:

3.2.1. Semantic Visual Embeddings (Pre-trained Deep Learning Models)

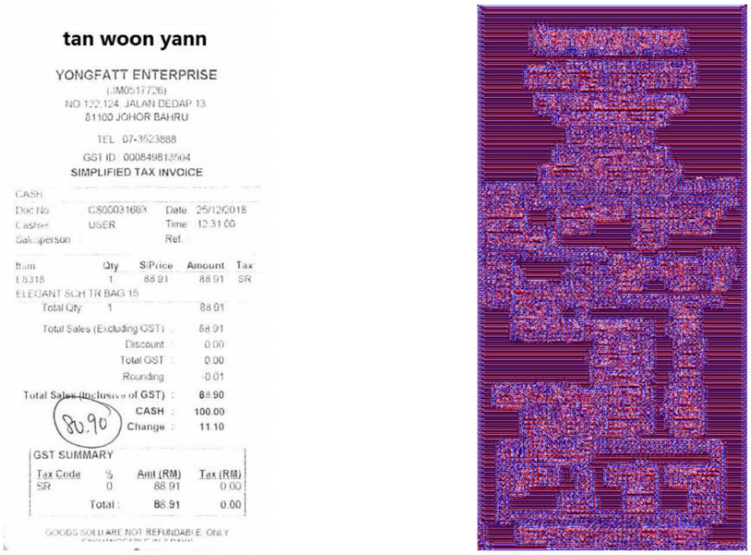
Two pre-trained convolutional neural networks (CNNs) were used to extract high-level semantic and structural features from receipt images:

- ResNet50, is a deep convolutional neural network (CNN) pre-trained on ImageNet that uses residual connections to learn hierarchical image features. When adapted to receipts, it encodes both local textures and global formatting. It is worth noting that ResNet50 is not trained explicitly for forgery detection, however, image tampering introduces anomalies such as pasted regions, inconsistent font rendering, or layout mismatches. Capturing the holistic structure of the receipt image allows for learnable changes in the image.

- Noiseprint, is also a CNN pre-trained model that extracts camera-specific noise patterns and suppresses semantic context. Digital forgeries typically affect native patterns, especially when conducting copy-move, splicing, or retouching. This results in a residual image that contains inconsistencies. By gathering the Noiseprints as statistical features, in hopes to detect unnatural noise signatures.

3.2.2. Sensor Noise Residual Analysis

This category also includes Noiseprint, which also serves to detect discrepancies in the low-level sensor noise. These residuals reflect intrinsic properties of the camera and can reveal tampering that may not be evident in the semantic content of the image. An example of Noiseprint map for receipt 'X00016469622.png' is as shown in Figure 5.



tan woon yann

YONGFATT ENTERPRISE
(JM0517726)
NO.122,124. JALAN DEDAP 13
81100 JOHOR BAHRU

TEL : 07-3523888

GST ID : 000849813504
SIMPLIFIED TAX INVOICE

CASH

Doc No	CS00031663	Date	25/12/2018
Cashier	USER	Time	12:31:00
Salesperson		Ref.	

Item	Qty	S/Price	Amount	Tax
E8318	1	88.91	88.91	SR
ELEGANT SCH TR BAG 15				
Total Qty	1		88.91	

Total Sales (Excluding GST) :	88.91
Discount :	0.00
Total GST :	0.00
Rounding :	-0.01
Total Sales (Inclusive of GST) :	**88.90**
CASH :	**100.00**
Change :	**11.10**

80.90

GST SUMMARY

Tax Code	%	Amt (RM)	Tax (RM)
SR	0	88.91	0.00
	Total :	88.91	0.00

GOODS SOLD ARE NOT REFUNDABLE, ONLY

Figure 5: Noiseprint map example for 'X00016469622.png'

3.2.3. Hand-Engineered Visual Features (Steganalysis-Inspired Methods)

The final image feature extraction approach involved a suite of manually designed statistical feature extractors inspired by steganalysis—originally developed to detect hidden information in images. These features are not learned from data but rather crafted based on digital forensic principles. It is noteworthy that these hand-engineered visual features offer a forensic lens into statistical irregularities and are complementary to the learned representations from deep models. Three open-source extractors were implemented:

- SPAM686, is a general-purpose steganalysis feature set that computes co-occurrence matrices over first-order spatial residuals in multiple directions. It captures statistical inconsistencies such as unnatural texture patterns or repeated structures, common in tampered images.

- s5x5_spam14hv_q1, is a specialized variant of the SPAM extractor, this method focuses on horizontal and vertical residuals within a 5×5 neighborhood. It is sensitive to directional anomalies, particularly those introduced during line-level edits or region pasting.

- s5x5_minmax22v_q1, is a second-order extractor that computes the minimum and maximum differences between vertically adjacent pixels. This method is designed to detect unnatural vertical transitions, such as those created by retouching, digital erasure, or misaligned overlays.

3.2.4. Dimensionality Reduction

To enhance model efficiency and reduce the risk of overfitting, dimensionality reduction was applied to several high-dimensional feature sets using Principal Component Analysis (PCA). The objective was to retain the most informative components while removing noise and redundancy from the data. Standardization was performed prior to PCA where applicable.

- ResNet50, reduced from 2048 original dimensions to 150 principal components after standardization.
- Noiseprint originally contained 144 dimensions; no dimensionality reduction was applied, but the features were standardized.
- SPAM686 was reduced from 686 dimensions to 150 principal components after standardization.
- s5x5_spam14hv_q1, Reduced from 338 dimensions to 150 principal components after standardization.
- s5x5_minmax22v_q1, reduced from 325 dimensions to 150 principal components after standardization.

3.3. Text Feature Extraction

3.3.1. Text-feature extraction with Hand crafted approach

The handcrafted feature extraction approach was designed to capture both structural and semantic characteristics of OCR-transcribed receipt text to aid in forgery detection. Receipt text files were aligned with image metadata and loaded into structured data frames for processing.

A variety of linguistic and formatting-based features were engineered to reflect common traits of authentic versus forged receipts, which consist of (1) textual structure metrics such as total character count, number of lines, word count, and average word length. (2) Character composition: counts of digits, uppercase/lowercase letters, and special characters. (3) Receipt-specific patterns: detection of date-like formats and numeric sequences typically found on real receipts.

To supplement these handcrafted indicators with semantic information, TF-IDF vectors (using unigrams and bigrams) were computed based on the top 100 most informative terms. These vectors helped capture subtle word usage and phrasing differences that might signal falsification.

The handcrafted and TF-IDF features were concatenated to form a unified feature vector. All features were standardized to ensure scale consistency before being passed to traditional classifiers. This composite representation enabled the models to detect both superficial formatting anomalies and deeper semantic inconsistencies indicative of document tampering.

3.3.2. Text-feature extraction with pre-trained models (DistilBERT)

To further enrich the feature space, a pretrained transformer model (DistilBERT) was used to extract contextual embeddings from the OCR-transcribed text. Each document was tokenized using the DistilBERT tokenizer, with sequences padded and truncated to a fixed length of 128 tokens. The model produced 768-dimensional embedding vectors that captured the semantic content of each receipt at a deep, contextual level. These embeddings encoded linguistic relationships and subtle variations in language that are often missed by surface-level features.

The model was fine-tuned for the binary classification task of forgery detection. To handle the substantial class imbalance (much few forged examples), a custom focal loss function was used to emphasize difficult or minority-class examples. Early stopping based on validation AUC was applied to prevent overfitting.

By leveraging pretrained language representations, this approach enabled the model to detect nuanced textual patterns and linguistic irregularities often associated with forged documents, providing a complementary and often more powerful signal than handcrafted features alone.

To further elaborate on DistilBERT, it is a compact and efficient transformer-based language model developed to address the computational limitations of larger models such as BERT (Bidirectional Encoder Representations from Transformers). While BERT achieved state-of-the-art performance across various natural language processing (NLP) tasks, its large size and slow inference time presented practical challenges for deployment in real-time and resource-constrained environments.

To mitigate these issues, DistilBERT was created using a technique known as knowledge distillation, wherein a smaller "student" model is trained to replicate the behavior of a larger "teacher" model, in this case, BERT. This approach allows DistilBERT to retain approximately 97% of BERT’s performance while being 60% faster and 40% smaller in terms of model size.

Despite its reduced complexity, DistilBERT effectively captures semantic and contextual information in text, making it well-suited for a range of downstream NLP applications such as text classification, sentiment analysis, and question answering. Its balance of performance and efficiency makes it a practical choice for both academic research and real-world deployment scenarios.

It is noteworthy although the original BERT model with different modifications were tested on this dataset, it was ultimately deemed unsuitable due to the dataset’s relatively small size and significant class imbalance. In contrast, DistilBERT provided a more computationally efficient and robust alternative for this specific dataset as a feature extraction algorithm.

3.4. Classifier Modeling: Image and Text Modalities

To evaluate the efficacy of various feature representations for receipt forgery detection, the image and text data streams were modeled independently within a unified methodological framework. This strategy enabled modular experimentation, allowing for a fair comparison between the two feature sets and supporting the development of final models that integrate both visual and textual modalities. For each feature set, the extracted representations were aligned with metadata labels, and samples with known annotation inconsistencies were excluded from the evaluation set.

All features were standardized to have zero mean and unit variance prior to model training. When appropriate, Principal Component Analysis (PCA) was applied to reduce dimensionality while preserving the most discriminative components. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied separately to each feature set, augmenting the minority class (forged receipts) within the training data to promote balanced learning during classifier training. A suite of widely used supervised learning models was applied independently to both image and text features:

- Support Vector Machines (SVM): linear and radial basis function kernels
- Logistic Regression
- Random Forests
- Multi-layer Perception (MLP) Neural Network (NN)
- XGBoost

Models were trained on the augmented training data and evaluated on a test set containing only receipts that appeared in both feature representations. To leverage the complementary strengths of the two text representations, two different types of ensemble learning were employed:

1. Stacking Ensemble: For each base classifier, a 5-fold stratified stacking approach was performed. First-layer base models were trained separately on different feature types, such as ResNet50/SPAM686 for image or handcrafted/DistilBERT for text, and their out-of-fold predicted probabilities were used as inputs to a second layer meta-classifier of the same type.

2. Probability Averaging Ensemble: Alternatively, the output probabilities from models trained on different features were averaged to form final predictions. The simpler ensemble strategy followed a more majority voting methodology that combined the strengths across the independent models

All models and ensembles were assessed using standard classification metrics including:

- Accuracy
- ROC AUC
- Confusion Matrix
- Precision, Recall, and F1-score

These evaluations were conducted consistently across all steps in the modeling pipeline, enabling a direct comparison on how visual and textual information contribute to forgery detection.

3.5. Class Imbalance Handling

To address class imbalance, all training sets were balanced using SMOTE (Synthetic Minority Oversampling Technique). SMOTE was applied separately to each feature set for both text and image respectively, to oversample the minority class (forged receipts) and ensure a balanced distribution before training the respective classifiers.

3.6. Late Fusion (Hybrid multimodal modeling for Image-Text)

To leverage the complementary strengths of image and text features, a late fusion strategy was employed by combining the predicted probabilities from the best-performing models. Specifically, a weighted sum of the output probabilities was computed from a Logistic Regression model trained based on image features and a Neural Network trained based on text features. This fusion approach enabled the integration of distinct signals from both modalities, leading to improved overall classification performance, particularly emphasis in detecting a higher number of fraudulent cases while achieving a more balanced trade-off between precision and recall.

The fused probability was calculated based on the following equation:

$$\boldsymbol{fused_{prob}} = (\boldsymbol{W} \times \boldsymbol{image_{prob}}) + ((\boldsymbol{1} - \boldsymbol{W}) \times \boldsymbol{text_{prob}})$$

where $\boldsymbol{W} \in [\boldsymbol{0}, \boldsymbol{1}]$, is the weight of the image model in the final prediction.

Threshold *T* was then applied to the fused probability to generate the final binary classification (1 for fraud and 0 for not fraud). Subsequently, a grid search over different values of *W* and *T* to determine the optimal fusion configuration was performed as well.

4. Evaluation and Final Results

4.1. Model Selection for Text-only models

Table 1: Consolidation of Text-only model performance

No.	Model	Ensemble Type	Accuracy	ROC AUC	Precision (Class =1)	Recall (Class =1)	F1-score (Class =1)
1	Linear SVM	Meta-classifier	0.648	0.578	0.161	0.290	0.207
2	Linear SVM	Prob. Averaging	0.628	0.577	0.150	0.290	0.198
3	Kernel SVM	Meta-classifier	0.709	0.455	0.094	0.097	0.095
4	Kernel SVM	Prob. Averaging	0.704	0.593	0.091	0.097	0.094
5	Logistic Regression	Meta-classifier	0.689	0.580	0.174	0.258	0.208
6	Logistic Regression	Prob. Averaging	0.689	0.562	0.200	0.323	0.247
7	Random Forest	Meta-classifier	0.796	0.517	0.091	0.032	0.048
8	Random Forest	Prob. Averaging	0.806	0.564	0.267	0.129	0.174
9	**MLP Neural Network**	**Meta-classifier**	**0.816**	**0.687**	**0.333**	**0.161**	**0.217**
10	**MLP Neural Network**	**Prob. Averaging**	**0.796**	**0.685**	**0.320**	**0.258**	**0.286**
11	XGBoost	Meta-classifier	0.796	0.533	0.154	0.065	0.091
12	XGBoost	Prob. Averaging	0.801	0.566	0.214	0.097	0.133

With respect to Table 1, based on all the text-only models evaluated, Model 9 (MLP Meta-classifier) and Model 10 (MLP Probability Averaging) emerged as the most viable choices for detecting forged receipts (class = 1). Model 9 achieved the highest ROC AUC (0.687) and highest overall accuracy (0.816), indicating strong discriminatory power. Model 9 has a precision of 0.333, which suggests that when a forgery is predicted, it is likely correct, which makes it suitable when false positives are costly.

In contrast, Model 10 demonstrated the highest F1-score (0.286) and higher recall (0.258) when compared to Model 9, indicating a better balance between capturing actual forgeries and minimizing false detection. With relatively decent precision (0.320) and AUC (0.685), it offers robust, balanced performance for scenarios prioritizing the detection of as many forgeries as possible. Therefore, depending on the operational objective, whether maximizing precision or improving recall, either model provides a viable solution.

It is noteworthy that all text classifiers were evaluated on the same 196 test samples, with 22 samples (1 forged, 21 non-forged) excluded due to missing predictions from the image classifier, which failed to converge on those specific instances. This is to standardize models before moving forward with multimodal hybrid models.

4.2. Model Selection for Image-only models

Table 2: Consolidation of Image-only model performance

No.	Model	Ensemble Type	Accuracy	ROC AUC	Precision (Class =1)	Recall (Class =1)	F1-score (Class =1)
1	Linear SVM	Meta-classifier	0.684	0.583	0.184	0.290	0.225
2	Linear SVM	Prob. Averaging	0.709	0.583	0.191	0.258	0.219
3	Kernel SVM	Meta-classifier	0.730	0.598	0.156	0.161	0.159
4	Kernel SVM	Prob. Averaging	0.714	0.572	0.195	0.258	0.222
5	Logistic Regression	Meta-classifier	0.719	0.600	0.239	0.355	0.286
6	**Logistic Regression**	**Prob. Averaging**	**0.730**	**0.598**	**0.280**	**0.452**	**0.346**
7	Random Forest	Meta-classifier	0.827	0.421	0.000	0.000	0.000
8	Random Forest	Prob. Averaging	0.832	0.602	0.000	0.000	0.000
9	MLP Neural Network	Meta-classifier	0.791	0.586	0.143	0.065	0.089
10	MLP Neural Network	Prob. Averaging	0.791	0.565	0.188	0.097	0.128
11	XGBoost	Meta-classifier	0.816	0.416	0.000	0.000	0.000
12	XGBoost	Prob. Averaging	0.832	0.614	0.000	0.000	0.000

With respect to Table 2, based on all the image-only models evaluated, Model 6 (Logistic Regression with Probability Averaging) demonstrated the best overall performance in detecting forgeries (class = 1). It achieved the highest F1-score (0.346), highest recall (0.452), and strong precision (0.280), indicating that it captured nearly half of the true forged instances while maintaining a reasonable false positive rate. It was able to achieve a ROC AUC of 0.598, which is also ranked among the top performers, suggesting good ranking ability despite class imbalance.

In addition, Model 5 (Logistic Regression Meta-classifier) with a precision of 0.239, recall of 0.355, and an F1-score of 0.286, which has also performed well, by offering a slightly more conservative balance between false positives and false negatives.

On the other hand, although Random Forest and XGBoost models achieved the highest accuracy of approximately 0.83, their precision, recall, and F1-scores were all zero, indicating a complete failure to identify any forged receipts. Therefore, based on this set of image-only model performance data, it demonstrated that Logistic Regression models, in particular Model 6, is the most viable solution for image-only forgery detection.

It is noteworthy that all text classifiers were evaluated on the same 196 test samples, with 22 samples (1 forged, 21 non-forged) excluded due to missing predictions from the image classifier, which failed to converge on those specific instances. This is to standardize models before moving forward with multimodal hybrid models.

4.3. Feature set Contribution Analysis

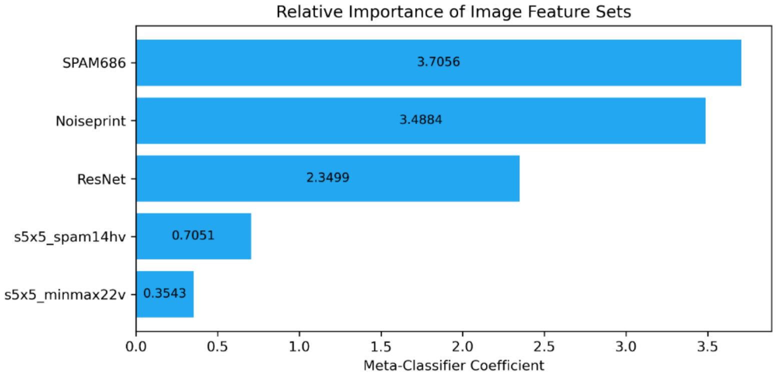


Figure 6: Relative Importance of Image Feature Sets

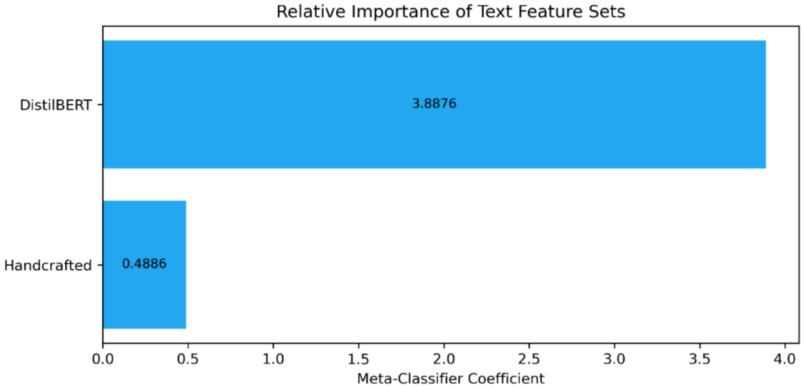


Figure 7: Relative Importance of Test Feature Sets

To understand the relative contributions of text and image modalities prior to late fusion modelling, a stacking-based meta-classifier was constructed and analyzed. This feature set contribution analysis aimed to interpret how each feature set influenced the final prediction by examining the learned coefficients of the meta-classifier, as depicted in Figure 6 and 7.

The stacking ensemble was trained using five-fold cross-validation on the training set, with a MLP neural network serving as the base model for text features and logistic regression for image features. In this context, the magnitude of each coefficient serves as an indicator of the corresponding modality's impact on the ensemble's decision-making process, where larger absolute values suggest higher relevance, which serves as a foundation for the subsequent late fusion strategies, which combine model predictions rather than raw features.

4.4. Model selection for Late Fusion Models (Image-and-Text)

Two combinations of text-and-image models were evaluated, and the top 10 configurations with class 1 recall ≥ 0.3 were selected for each model.

4.4.1. Late Fusion Model 1: Probabilistic Averaging of Logistic Regression (Image) and MLP Neural Network Meta-Classified (Text)

Table 3: Consolidation of Late Fusion Model 1 performance

No.	Weigh (W)	Threshold(T)	Recall_0	Precision_0	Recall_1	Precision_1	Accuracy	AUC	F1_1
1	0.250	0.200	0.818	0.894	0.484	0.333	0.765	0.697	0.395
2	0.100	0.150	0.818	0.888	0.452	0.318	0.760	0.702	0.373
3	0.150	0.200	0.867	0.872	0.323	0.312	0.781	0.700	0.317
4	0.300	0.200	0.770	0.901	0.548	0.309	0.735	0.696	0.395
5	0.150	0.150	0.782	0.896	0.516	0.308	0.740	0.700	0.386
6	0.000	0.150	0.848	0.875	0.355	0.306	0.770	0.687	0.328
7	0.200	0.200	0.848	0.875	0.355	0.306	0.770	0.699	0.328
8	0.650	0.400	0.861	0.871	0.323	0.303	0.776	0.652	0.312
9	0.700	0.400	0.830	0.878	0.387	0.300	0.760	0.650	0.338
10	0.050	0.150	0.842	0.874	0.355	0.297	0.765	0.700	0.324

The Late Fusion Model 1 is a combination of image-only Model 6 (Logistic Regression with Probability Averaging) and text-only Model 9 (MLP Meta-classifier).

With respect to Table 3, it can be observed that for Late Fusion Model 1, class 1 recall (recall_1) reaches its highest value around W = 0.3 (0.548), with relatively strong performance in the range W = 0.1–0.3. Beyond this range, recall_1 begins to decline, reaching 0.355 at W = 0.65. Class 1 precision (precision_1), however, remains consistently low across all W values and does not show meaningful improvement at higher weights. For Model 1, moderate fusion weights (W = 0.25–0.3) provide the most effective trade-off for fraud detection, particularly when prioritizing class 1 recall.

4.4.2. Late Fusion Model 2: Probabilistic Averaging of Logistic Regression (Image) and MLP Neural Network Probability Averaging (Text)

Table 4: Consolidation of Late Fusion Model 2 performance

No.	Weight (W)	Threshold(T)	Recall_0	Precision_0	Recall_1	Precision_1	Accuracy	AUC	F1_1
1	0.400	0.450	0.861	0.882	0.387	0.343	0.786	0.709	0.364
2	0.450	0.450	0.861	0.882	0.387	0.343	0.786	0.710	0.364
3	0.850	0.450	0.855	0.881	0.387	0.333	0.781	0.706	0.358
4	0.350	0.450	0.855	0.881	0.387	0.333	0.781	0.710	0.358
5	0.500	0.450	0.855	0.881	0.387	0.333	0.781	0.711	0.358
6	0.600	0.450	0.842	0.885	0.419	0.333	0.776	0.700	0.371
7	0.150	0.450	0.848	0.881	0.387	0.324	0.776	0.704	0.353
8	0.300	0.450	0.855	0.876	0.355	0.314	0.776	0.709	0.333
9	0.650	0.450	0.818	0.882	0.419	0.302	0.755	0.693	0.351
10	0.100	0.450	0.830	0.878	0.387	0.300	0.760	0.700	0.338

The Late Fusion Model 2 is a combination of image-only Model 6 (Logistic Regression with Probability Averaging) and text-only Model 10 (MLP Neural Network Probability Averaging).

With respect to Table 4, it can be observed that for Late Fusion Model 2, class 1 recall (recall_1) remains mostly stable at 0.387 across W = 0.1–0.5. In contrast, at W = 0.6–0.65, recall_1 increases slightly to 0.419. However, this improvement comes with a slight drop for class 1 precision (precision$_1$). Overall, Model 2 benefits modestly from higher W values (W = 0.6–0.65) in terms of recall$_1$, although precision$_1$ slightly declines.

4.4.3. Optimization for Late Fusion Models

Across both models, the classification threshold **T** also affects the balance between class 1 recall (recall$_1$) and precision (precision$_1$):

- **Lower thresholds** (T < 0.5) typically improve **class 1 recall (recall$_1$)**, capturing more positive cases but often at the expense of precision.
- **Higher thresholds** (T >= 0.5) increase **class 1 precision** (**precision$_1$)** by reducing false positives but tend to reduce class 1 recall.

A lower classification threshold is preferable when maximizing class 1 recall is the priority, particularly in fraud detection tasks where missing a positive case is likely costlier than a false alarm.

The optimal configurations for each fusion pipeline demonstrated a strong balance between forgery detection (class 1 recall) and false positive control (class 1 precision), are as follows:

- **Late Fusion Model 1: Probabilistic Averaging of Logistic Regression (Image) and MLP Neural Network Meta-Classified (Text) tuned to *W = 0.25, T = 0.20***

With respect to Late Fusion Model 1 configuration with W = 0.25, T = 0.20, it was able to achieve solid balance between fraud detection (class 1 recall) and false positive control (class 1 precision), as shown in Table 5.

Table 5: Classification Report for Late Fusion Model 1 (W = 0.25, T = 0.20)

Classification Report for Late Fusion Model 1 (*W = 0.25, T = 0.20)*				
Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.894	0.818	0.854	165
1	0.333	0.484	0.395	31
Metric	**Precision**	**Recall**	**F1-Score**	**Support**
Accuracy	N.A.	N.A.	0.765	196
Macro Average	0.614	0.651	0.625	196
Weighted Average	0.805	0.765	0.782	196

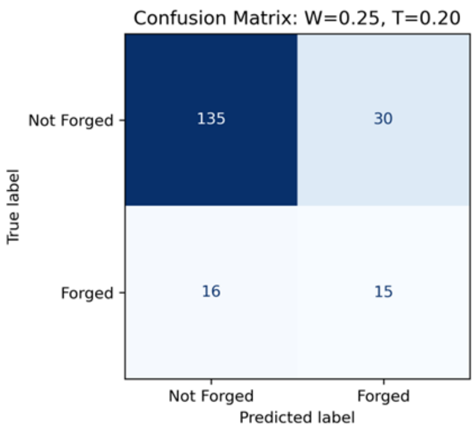


Figure 8: Confusion Matrix for Late Fusion Model 1 (W = 0.25, T = 0.20)

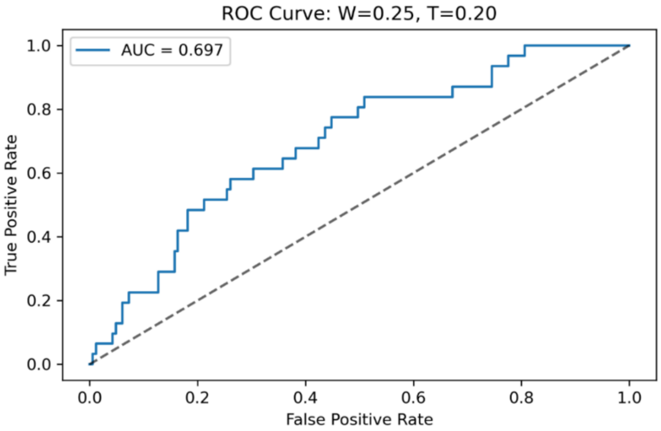


Figure 9: ROC Curve for Late Fusion Model 1 (W = 0.25, T = 0.20)

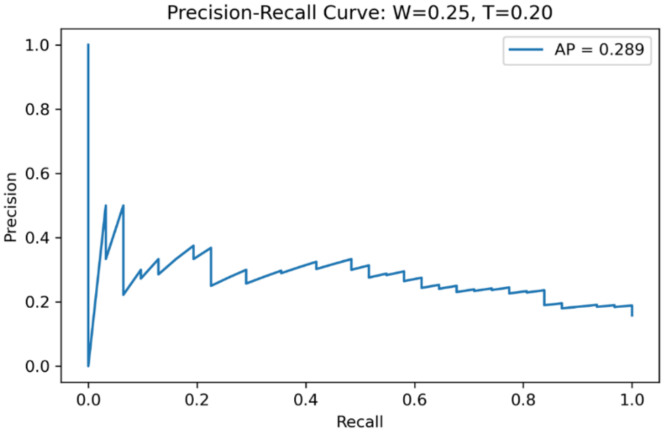


Figure 10: Precision-Recall Curve for Late Fusion Model 1 (W = 0.25, T = 0.20)

- **Late Fusion Model 2: Probabilistic Averaging of Logistic Regression (Image) and MLP Neural Network Probability Averaging (Text) tuned to *W = 0.60, T = 0.45***

With respect to Late Fusion Model 2 configured with W = 0.60, T = 0.45, it demonstrated slightly higher overall accuracy (0.776) and AUC (0.709) when compared to Late Fusion Model 1, indicating better overall ranking performance. However, it achieved a lower class 1 recall of 0.419, suggesting reduced sensitivity to forged cases. This configuration resulted in fewer false positives and may be more suitable for applications where false alarms are costlier than missed detections.

Table 6: Classification Report for Late Fusion Model 2 (W = 0.60, T = 0.45)

Classification Report for Late Fusion Model 2 (*W = 0.60, T = 0.45)*				
Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.885	0.842	0.863	165
1	0.333	0.419	0.371	31
Metric	Precision	Recall	F1-Score	Support
Accuracy	N.A.	N.A.	0.776	196
Macro Average	0.609	0.631	0.617	196
Weighted Average	0.798	0.776	0.786	196

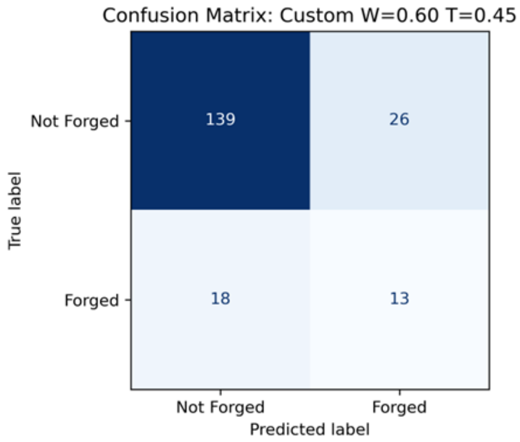


Figure 11: Confusion Matrix for Late Fusion Model 2 (W = 0.60, T = 0.45)

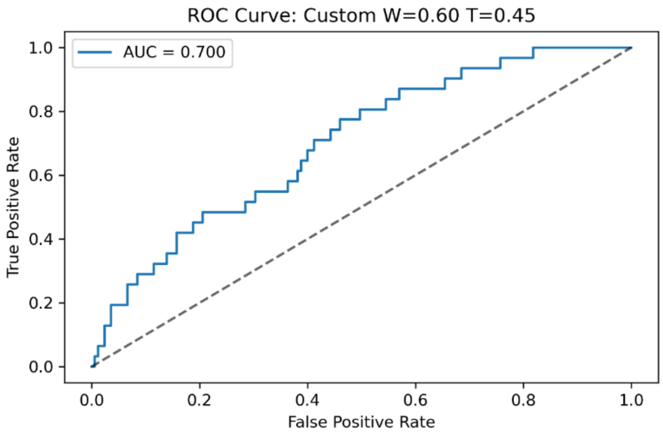


Figure 12: ROC Curve for Late Fusion Model 2 (W = 0.60, T = 0.45)

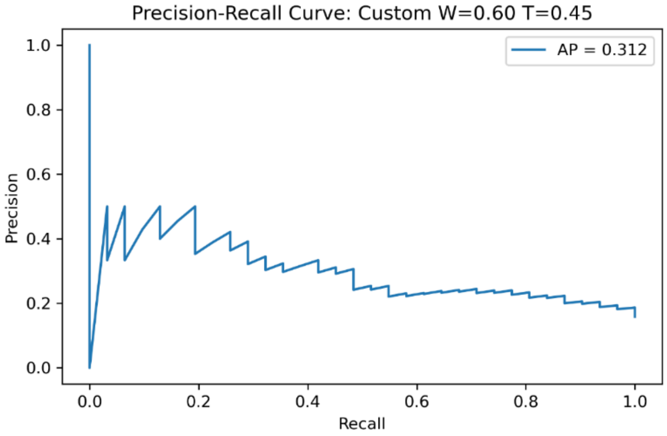


Figure 13: Precision-Recall Curve for Late Fusion Model 2 (W = 0.60, T = 0.45)

4.5. Summary for Best-performing models across all three categories (Text-only, Image-only, and Late Fusion)

The following selection was based on maximizing class 1 recall (forgery detection), with acceptable precision and overall performance. It is also noteworthy that Late Fusion Model 2 should also be considered if a different selection criteria was utilized, as it demonstrated benefits in other model evaluation metric.

Table 7: Summary table of Best-performing models by Modality

Model Type	**Model Description**	**Accuracy**	**Recall (Class 1)**	**Precision (Class 1)**	**F1 (Class 1)**	**ROC AUC**	**Notes**
Text-only (Model 10)	MLP Neural Network (Prob. Averaging)	0.796	0.258	0.320	0.286	0.685	Best text-only model in recall and F1
Image-only (Model 6)	Logistic Regression (Prob. Averaging)	0.730	0.452	0.280	0.346	0.598	Best image-only model in recall and F1
Late Fusion Model 1	LR (Image) + MLP NN Meta (Text), W = 0.25, T = 0.20	0.765	**0.484**	0.333	**0.395**	0.697	Best recall and F1 overall, strong balance
Late Fusion Model 2	LR (Image) + MLP NN Prob. (Text), W = 0.60, T = 0.45	0.776	0.419	0.333	0.371	0.709	Highest accuracy and AUC, lower sensitivity

5. Conclusion

This project addressed the growing challenge of receipt forgery detection by developing a multimodal machine learning framework that prioritizes both accuracy and interpretability as key requirements in regulated domains such as finance, insurance, and auditing. By leveraging both image and text modalities, the system integrated pretrained embeddings (ResNet50, Noiseprint, DistilBERT), interpretable handcrafted features, and ensemble learning to robustly classify receipts as authentic or forged.

Despite challenges including severe class imbalance and a limited number of forged samples, the framework proved effective. A comprehensive comparison across text-only, image-only, and late fusion models reveals the strengths and limitations of each modality. Among text-only models, the MLP with probability averaging (Text-only Model 10) yielded the most balanced performance, with an accuracy of 0.796, class 1 recall of 0.258, and F1-score of 0.286, reflecting moderate success in capturing forged receipts. In contrast, image-only models demonstrated stronger forgery recall, with Logistic Regression (Image-only Model 6) achieving recall of 0.452 and F1-score of 0.346, indicating that visual patterns provided more direct clues for forgery detection.

However, the greatest performance gains were observed in the late fusion models, which combined both text and image features. Specifically, Late Fusion Model 1 (W = 0.25, T = 0.20) achieved the highest class 1 recall of 0.484, F1-score of 0.395, and macro F1 of 0.625, demonstrating superior ability to detect forged receipts while maintaining reasonable precision (0.333) and overall accuracy (0.765). Though Late Fusion Model 2 (W = 0.60, T = 0.45) slightly outperformed in overall accuracy (0.776) and AUC (0.709), it traded off recall (0.419) and thus was less sensitive to fraud cases. This comparison confirms that a multimodal fusion strategy, particularly Late Fusion Model 1, offers the most effective and balanced approach for reliable forgery detection.

While the lack of “before-and-after” forgery pairs limited direct analysis of tampering artifacts, and some handcrafted visual features were excluded due to low discriminative power or computational cost, this limitation realistically reflects real-world scenarios, where original (pre-forgery) documents are typically unavailable. Despite these constraints, the final pipeline remains modular, scalable, and interpretable. Therefore, this project is considered relatively successful in addressing forgery detection.

Reference

- 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). https://ieeexplore.ieee.org/document/7780459
- Cozzolino, D., & Verdoliva, L. (2019). *Noiseprint: A CNN-based camera model fingerprint*. *IEEE Transactions on Information Forensics and Security*, 15, 144–159. https://ieeexplore.ieee.org/document/8713484
- Pevný, T., Bas, P., & Fridrich, J. (2010). *Steganalysis by subtractive pixel adjacency matrix*. *IEEE Transactions on Information Forensics and Security*, 5(2), 215–224. https://ieeexplore.ieee.org/document/5437325
- Fridrich, J., & Kodovský, J. (2012). *Rich models for steganalysis of digital images*. *IEEE Transactions on Information Forensics and Security*, 7(3), 868–882. https://ieeexplore.ieee.org/document/6197267
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. *arXiv preprint arXiv:1910.01108*. https://arxiv.org/abs/1910.01108
- FIND-IT Fraud Detection Contest. (2023). *Receipt Fraud Detection Challenge Report*. https://hal.science/hal-02316399/file/fraud-detection-contest-report.pdf
- GRIP-Unina. (n.d.). *Noiseprint Source Code*. GitHub. https://github.com/grip-unina/noiseprint
- Binghamton University. (n.d.). *Feature Extractors for Steganalysis*. Digital Data Embedding Lab. https://dde.binghamton.edu/download/feature_extractors/