j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

Image Document Classification Prediction based on SVM and gradient-boosting Algorithms

[1*]Ahmed H. Salman, [2] Waleed A, Mahmoud Al-Jawher

[1]*Iraqi Commission for Computers and Informatics, Informatics Institute of Postgraduate Studies, Baghdad, Iraq.*

[2]*College of Engineering, Uruk University, Baghdad, Iraq.*

phd202110677@iips.edu.iq

Abstract Image document classification is crucial in various domains, including healthcare, finance, and security. Automatically categorizing images into predefined classes can significantly improve data management and decision-making processes. For this research, we investigate the effectiveness of two machine learning algorithms, Support Vector Machines (SVM) and Gradient Boosting, for image document classification. First, we preprocess the image data by extracting relevant features, such as Image Embedding, to create a feature vector for each image. These features are essential for representing the content of the images accurately. Next, we apply SVM, a robust supervised learning algorithm, to train a classification model. SVM aims to Determine the optimal hyperplane for effectively distinguishing the images into different classes while maximizing the margin. Furthermore, we explore the Gradient Boosting algorithm, an ensemble learning method combining multiple weak learners to create a robust classifier. We experimented with different classification results with ten classes. We employ Multiple measures, including accuracy, precision, recall, F1-score, and ROC-AUC, are used to assess the performance of the SVM and Gradient Boosting models. The higher result of 0.964 for SVM compared with Adaboost is achieved. 0.853.

Crossref iob 10.36371/port.2023.4.5

Keywords: *Document Classification, Feature Extraction, Gradient Boosting algorithm, and Support Vector Machines (SVM).*

1. INTRODUCTION

All corporate Communication and record-keeping depend on documents [1]. Automatic document information extraction is difficult [2]. Typically, Prior to information extraction, physical documents undergo the process of being scanned and photographed. Many Document Image Processing Pipelines require document classification. Document classification improves document processing systems [3]. Thus, several document classification methods leverage text content [5], document structure [7] or both [12]. This field has advanced, especially with deep learning [14]. Since AlexNet [16], deep neural network research and performance have grown significantly. Numerous experiments and studies have been undertaken with computer vision models, including VGNet [17], Resnet [18], and InceptionNet [19]. The models performed well with classification tasks since the documents are viewed as images [27]. Many documents are essentially identical but have different text. Thus, these document images communicate high-level structural information, but low-level traits that can distinguish visually comparable images have been neglected for a long time [28]. Several articles [13] consider adding features to increase accuracy. These papers yielded cutting-edge outcomes. They employ a robust OCR engine, such as [20], which is used to extract text from document images and acquire knowledge of both textual and visual characteristics. This helps solve visually similar document instances. We use SVM and AdaBoost for Image Document Classification Prediction in this paper. The structure of this paper is as follows: Part 2 will focus on the classification of related work. Section 3 illustrates the proposed methodology. Section 4 provides an explanation of the results and a discussion of classification. Finally, Section 5 delves into the conclusion and prospects discussion

2. RELATED WORK

Much work has been completed on the classification of document images. Prior research mostly concentrated on extracting textual characteristics from the documents, leading to its widespread recognition as text document categorization [24]. Using SVM and AdaBoost Algorithms for the document images in our proposed paper. In this related work, the authors of Afzal et al.'s paper[1], The utilization of Alexnet as a classifier model, considering documents as images, has facilitated further investigation into the inclusion of visual features. This study demonstrated that the utilization of transfer learning significantly enhanced the outcomes. Furthermore, the dataset was utilized for training the model

Ahmed H. Salman, Waleed A, Mahmoud Al-Jawher. Image Document Classification Prediction based on SVM and gradient-boosting Algorithms. *Journal port Science Research*, 6(4), pp. 348-356. https://doi.org/10.36371/port.2023.4.5

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

by transfer learning. Subsequently, this model was employed to categorize the Tobacco-3482 dataset. Harley et al. [8] created a vast RVL-CDIP dataset comprising 400,000 picture documents. Abuelwafa et al. [25] proposed an unsupervised categorization methodology. The author explicitly states that the model was exclusively trained using the input image and did not utilize any annotated data. The purpose of the model was to analyze the visual characteristics of the input image. The authors trained the model [26] on an auxiliary task where the input was linked to a distinct label and expanded to several images using a data augmentation technique. This strategy significantly enhanced the performance of the model. Roy et al. [26] employed generic, compact, and powerful CNNs in a separate investigation to analyze the characteristics of the input images.The authors in paper [26], Two Stream Deep Networks for Document Image Classificationutilizing Reuters and NLPCC 2014 news categorization datasets.

3. THE PROPOSED METHODOLOGY

This study aimed to determine the tobacco 3482 dataset, a representation that produces the best results, as different feature extraction strategies can impact classification performance in different ways. The study describes how orientation was categorized in the Tobacco 3482 dataset using image models and classifiers. SVM and AdaBoost Algorithm algorithm-optimized machine-learning techniques were employed to enhance image classification expertise. The proposed solution addresses these challenges by incorporating feature extraction and machine learning. Feature extraction is crucial in defining image content and extracting essential information from documents. Multiclass classification was performed using advanced machine learning methods such as AdaBoost and SVM. These algorithms utilized the extracted features to handle complex document classification tasks across multiple classes. Figure 1 provides a schematic representation of the research's model.

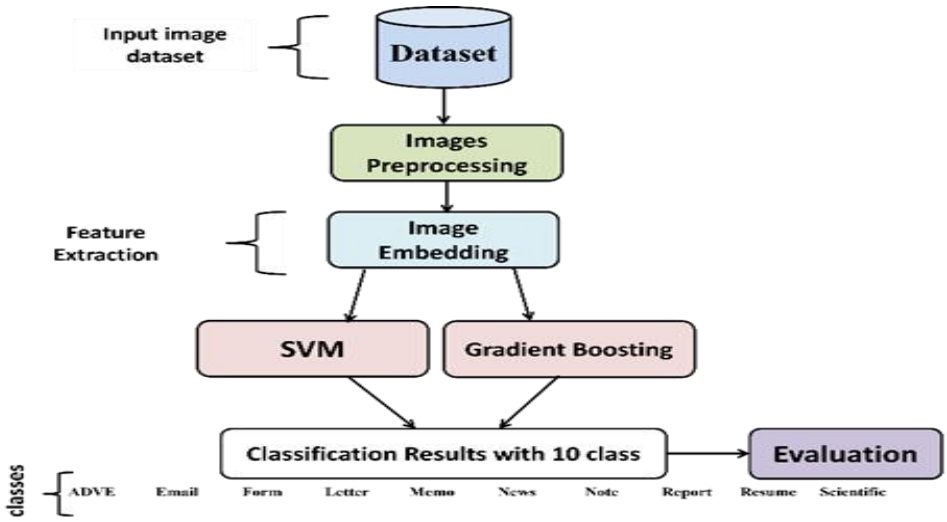

Fig. 1. Framework of multiclass Document Classification model.

It effectively conveys that the following subsections will delve into more detailed discussions of each phase.

1. Document Image Dataset: In the context of image documentation prediction, you would typically have a dataset of document images. These images can be of various types: memos, letters, reports, and scientific research papers. Or any other document category. This paper is based on the Tobacco 3482 dataset, a widely used resource in document analysis and optical character recognition (OCR). It was created to facilitate research and development in these areas. Contents and Size in Dataset as suggested by its name, the dataset contains 3,482 scanned document images. These images came from the legacy tobacco industry documents and were made publicly available following legal action against major tobacco companies. The dataset contains various document types, such as memos, letters, reports, and scientific research papers. This diversity makes it suitable for training and testing document analysis systems [10].

2. Processing Images: To use VGG-19 for feature extraction, you would follow these steps: Load the pre-trained VGG-19 model.

Remove the final classification layer, as you are interested in the features rather than class predictions.

Preprocess your document images to ensure they match the input requirements of the VGG-19 model (e.g., resizing to a specific input size and normalizing pixel values).

Pass each preprocessed image through the modified VGG-19 model to extract features. These extracted features are typically a high-dimensional vector.

3. Feature Extraction: When we talk about feature extraction or data embedding, we mean that we use the layers of the VGG-19 model to transform input images into a more

Ahmed H. Salman, Waleed A, Mahmoud Al-Jawher. Image Document Classification Prediction based on SVM and gradient-boosting Algorithms. *Journal port Science Research*, 6(4), pp. 348-356. https://doi.org/10.36371/port.2023.4.5

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

compact and meaningful representation. The lower layers of the model The lower levels of the model extract basic visual elements such as edges and textures, while the upper layers capture more advanced and intricate visual aspects such as forms and patterns.

4. Feature Vector Representation: Each document image is transformed into a feature vector where Every element in the vector signifies a certain characteristic or feature extracted by the VGG-19 layers. This feature vector captures the salient visual information from the document image.

5. Prediction:Model: This section describes the algorithm employed in the construction of a document categorization model.

A. AdaBoost Algorithm: AdaBoost, commonly called Adaptive Boosting, is a popular machine-learning approach for classification and regression. Ensemble learning combines outputs from multiple weak learners, such as decision trees or simple models, to create a robust learner. AdaBoost can adapt to the dataset, mitigate overfitting, and work well with diverse weak learners. However, this approach may be sensitive to noise and outliers. The user text contains no information that can be rephrased [19].

Support Vector Machines: The structural risk minimization principle serves as the foundation for SVM. The main goal is to increase classification accuracy by creating an ideal separation hyperplane. It is frequently used to solve binary classification issues.The sample set $S = X1, Y1), \ldots, (Xi, Yi), \ldots, (Xn, Yn)\}$ X is an input vector and Y is a classification result with a value of –1 or 1, is based on the assumption that there are n training samples. The best separating for linear separable problems is studied using the Support Vector Machine (SVM) model. $\phi(x)$ is considered one non-linear function for the non-linear separable problems, which aims to transfer the Map the input space to a feature space with a higher number of dimensions. Within the feature space of higher dimensions, we can derive the ideal separation plane, $\phi(x)\cdot w + b = 0$. The SVM model is concerned with $(xi\cdot xj)$, which is the inner product of the two vectors. As a result, all that is required of e is to compute the inner product of o on-linear functions $(\phi(xi)\cdot\phi(xj))$, where $K(xi \cdot xj) = \phi(xi) \cdot \phi(xj)$, $K(xi \cdot xj)$is known as the kernel function. There are numerous forms for the kernel function, including polynomial, RBF, and others. The kernel in the Support Vector Machine (SVM) is considered to be the Gaussian function exp $(-\frac{(x_1-x_2)^2}{\sigma^2})$, in this study.

Feature vectors are input to this section to a machine-learning model for document-type prediction. This model can learn to recognize patterns and relationships between the extracted features and the types of documents.

6. Evaluation: The evaluation of a classifier will be done using precision, recall, f-measure, and accuracy, sensitivity, and specificity are calculated as shown in (Eq. 10, Eq.11, Eq 12, Eq 13, Eq 14, Eq . 15)[23-25].

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (1)$$

The actual positive rate (T.P.) and false positive rate significantly impact positive instance recall or sensitivity.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

The following equation calculates accuracy, percentage of accurate predictions, and false-negative rate (F.N.).

$$\text{Accuracy} = (\text{TP} + \text{FP})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (3)$$

The term "T.N." means true negative, whereas "sensitivity" means the number of positive records that give the proper result.

$$\text{Sensitivity} = \text{TP}/\text{TP} + \text{FN} \quad (4)$$

Particularity is accurately arranging positive records from every positive paper.

$$\text{Specificity} = \text{TN}/\text{TN} + \text{FP} \quad (5)$$

The F-measure runs many data recovery accuracy norms and examines measurements.

$$\text{F1 Score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \quad (6)$$

Correct classification employs True Positives (T.P.) and False Positives (F.P.), while incorrect classification uses False Negatives (F.N.). A test's document classification accuracy is determined by its sensitivity and specificity. The ROC curve illustrates the trade-off between true and false positives. When the emphasis is skewed and false positives are ignored, the results are likely to reflect the accuracy of genuine positives primarily. Conversely, if true positives are neglected and false positives are emphasized, the scores will reflect recall. The Area Under the Curve (AUC) measures classifier efficiency[26]. Note that this classifier can be reinforcement by optimization method [31-33], where you train a learning agent to solve a complex problem by simply taking the best actions given a state, with the probability of taking each action at each state defined by a policy. An example is running a maze, where the position of each cell is the 'state', the 4 possible directions to move are the actions, and the probability of moving each direction, at each cell (state) forms the policy. This will be a topic of future work [34-36].

4. RESULTS AND DISCUSSION FOR CLASSIFICATION:

The classification results using Support Vector Machines (SVM) and Gradient Boosting on the Tobacco 3482 dataset. The accuracy, precision, recall, and F1 score metrics are provided. The result shown in Table 1.

Ahmed H. Salman, Waleed A, Mahmoud Al-Jawher. Image Document Classification Prediction based on SVM and gradient-boosting Algorithms. *Journal port Science Research*, 6(4), pp. 348-356. https://doi.org/10.36371/port.2023.4.5

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

Table I. *The Performance Of Various Document Classification Algorithms On The Tobacco Dataset.*

Model	AUC	CA	F1	PRECISION	RECALL
SVM	0.964	0.731	0.721	0.731	0.731
Adaboost	0.853	0.750	0.752	0.753	0.751

Table 2 is a confusion matrix of a classification model's predictions for Gradient Boosting. Here's a detailed explanation of the results:

ADVE (Advertisements): This class is correctly predicted 80.6% of the time, which is relatively high. It is sometimes confused with News (5.7%) and Note (3.7%).

Email: Correctly classified 78.6% of the time, with a small confusion with Memo (5.0%) and Form (2.1%).

Form: This class has a 65.4% correct prediction rate, but there's notable confusion with Resume (14.0%) and Letter (2.1%).

Letter: Correctly predicted 51.0% of the time, one of the lower correct prediction rates. There's a significant amount of confusion with memos (13.3%) and Emails (3.5%).

Memo: Has a 44.9% correct prediction rate, with confusion occurring with Letter (21.1%) and Report (16.4%).

News: This class has a high correct prediction rate of 71.1%, with some confusion with ADVE (6.8%) and Note (3.7%).

Note: Correctly classified 64.0% of the time, often confused with Memo (7.4%) and Email (3.3%).

Report: This has a correct prediction rate of 36.8%, which is lower than others, indicating significant confusion with other classes, notably Letter (10.5%) and Memo (7.9%).

Resume: It has a high correct prediction rate of 41.9%, but there is confusion with Form (14.0%) and Scientific (12.4%).

Scientific: This class has a correct prediction of 30.3%, the highest rate of accurate prediction in its row but relatively low overall. It is often confused with Resume (8.1%) and Report (11.5%).

The diagonal shows the percentages for correct classifications, ideally 100% for a perfect model. The off-diagonal cells show misclassifications, which ideally should be 0%. In this matrix, there's a variability in the model's ability to classify the different types of documents correctly. Some classes like 'ADVE' and 'News' are predicted with higher accuracy, while others like 'Report,' 'Resume', and 'Scientific' have lower accuracy and higher confusion than others.

The matrix helps in understanding which classes are confused by the model, indicating where improvements are needed. For example, the model seems to struggle with distinguishing 'Report,' 'Resume,' and 'Scientific' documents accurately, perhaps due to similarities in their features. The confusion matrix is critical for diagnosing classification models and guiding further refinement.

Table 2. *Confusion Matrix Of A Classification Model's Predictions For Gradient Boosting*

Prediction / Actual	ADVE	Email	Form	Letter	Memo	News	Note	Report	Resume	Scientific
ADVE	80.6 %	0.6 %	1.7 %	1.0 %	1.4 %	5.7 %	3.7 %	2.4 %	0.0 %	1.6 %
Email	0.5 %	78.6 %	2.1 %	3.3 %	5.0 %	0.0 %	3.7 %	4.0 %	1.2 %	7.4 %
Form	3.6 %	3.3 %	65.4 %	2.1 %	8.3 %	3.1 %	11.0 %	2.4 %	14.0 %	7.0 %
Letter	0.9 %	3.5 %	5.5 %	51.0 %	13.3 %	1.9 %	5.9 %	21.2 %	15.1 %	11.9 %
Memo	1.4 %	5.5 %	9.2 %	21.1 %	44.9 %	2.5 %	7.4 %	16.4 %	11.6 %	12.7 %
News	6.8 %	0.6 %	2.4 %	0.8 %	1.7 %	71.1 %	3.7 %	0.8 %	2.3 %	8.2 %
Note	5.9 %	3.3 %	5.0 %	1.4 %	4.1 %	1.9 %	64.0	1.2	1.2	5.7
Report	0 %	1.6 %	1.7 %	10.5 %	7.9 %	1.9 %	0.0	36.8	4.7	11.5
Resume	0 %	0.5 %	2.1 %	3.8 %	4.6 %	0.6	0.0	2.4	41.9	3.7
Scientific	0.5	2.5	5.0	4.9	8.7	11.3	0.7	12.4	8.1	30.3

While using Support Vector Machines (SVM), the result in Table 3 is a confusion matrix of a classification model's predictions for Support Vector Machines (SVM). Here's a detailed explanation of the results:

The classes along the top and the left side include ADVE, Email, Form, Letter, Memo, News, Note, Report, Resume, and Scientific. These represent the documents the model has been trained to identify.

The diagonal from the top left to the bottom right shows the percentage of instances for each class that were correctly identified (true positives). For example, the model correctly identified 90.0% of the actual emails as emails and 96.7% of the actual resumes as resumes, which have relatively high success rates.

The off-diagonal numbers represent misclassifications (errors). For example, 5.6% of the forms were incorrectly

Ahmed H. Salman, Waleed A, Mahmoud Al-Jawher. Image Document Classification Prediction based on SVM and gradient-boosting Algorithms. *Journal port Science Research*, 6(4), pp. 348-356. https://doi.org/10.36371/port.2023.4.5

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

classified as ADVE and 3.8% of emails were misclassified as memos.

The most accurately predicted class is 'Resume' with 96.7% accuracy, followed by 'Email' with 90.0%, and 'Form' with 81.2%. These high percentages indicate that the model is particularly good at classifying these types of documents.

The least accurately predicted classes are 'ADVE' and 'Scientific', with 7.2% and 62.1% of the predictions being correct, respectively. These lower percentages could indicate that these categories are more challenging for the model, possibly due to similarities with other classes or insufficient training data.

The model commonly confuses 'Note' with 'Letter', as indicated by the 25.4% misclassification rate.

A relatively small percentage of classifications are spread across unrelated categories, indicating that while the model does make mistakes, it does not often confuse completely dissimilar document types.

This confusion matrix is crucial for model evaluation as it provides insight into the overall accuracy and how the model performs in each class.

By the incorporation of Swin transformer with transformation techniques, the process of selecting, combining, generating or adapting several features to efficiently solve accuracy and computation time problems. One of the motivations for studying Swin transformer is to build systems which can handle classes of problems rather than solving just one problem [29-30].

Table 3. *Confusion Matrix Of A Classification Model's Predictions For Support Vector Machines (Svm)*

Prediction / Actual	ADVE	Email	Form	Letter	Memo	News	Note	Report	Resume	Scientific
ADVE	7.2 %	0.3 %	0.3 %	0.0 %	0.1 %	3.0 %	0.6 %	0.3 %	0.0 %	0.6 %
Email	0.0 %	90.0 %	2.3 %	2.4 %	1.6 %	0.0 %	3.8 %	2.9 %	0.0 %	0.6 %
Form	5.6 %	2.0 %	81.2 %	2.4 %	6.8 %	0.0 %	3.8 %	0.9 %	3.3 %	0.6 %
Letter	1.0 %	1.3 %	1.8 %	68.9 %	17.5 %	0.0 %	1.3 %	25.4 %	0.0 %	2.8 %
Memo	3.0 %	2.3 %	4.8 %	11.8 %	57.9 %	0.0 %	0.0 %	9.4 %	0.0 %	11.9 %
News	7.9 %	0.5 %	0.0 %	0.7 %	0.4 %	87.0 %	1.3 %	0.0 %	0.0 %	3.4 %
Note	7.0 %	1.1 %	3.0 %	1.1 %	1.5 %	0.0 %	87.8 %	0.9 %	0.0 %	2.3 %
Report	1.3 %	1.8 %	1.0 %	8.8 %	3.2 %	0.6 %	0.0 %	45.7 %	0.0 %	13.6 %
Resume	0.3 %	0.2 %	0.8 %	1.5 %	3.7 %	0.0%	0.0 %	4.4 %	96.7 %	2.3 %
Scientific	1.7 %	0.5 %	5.0 %	2.4 %	7.4 %	9.5 %	1.3 %	10.0 %	0.0 %	62.1 %

Figure 3 depicts the Receiver Operating Characteristic (ROC) curve which shows the prediction model's performance across different categorization criteria after using Support Vector Machines (SVM), and Gradient Boosting. ROC curve comparisons between multiple classifiers are common, it reaches SVM (Support Vector Machine, a discriminative classifier officially defined by a separating hyperplane) and Gradient Boosting (a machine learning technique that develops a prediction model in an ensemble of weak prediction models, often decision trees).

The gradient-boosting model has very little advantage at specific thresholds, but otherwise, both models are almost overlapping and seem to have extremely comparable performance characteristics in the Map you gave. High performance is suggested by the fact that both curves are very near to the plot's upper-left corner. A flawless classifier is shown if the curves reach the upper left corner (0,1); a random guess is indicated if the curves follow the 45-degree line .For the majority of threshold values, both models have a high TPR and a low FPR, indicating that they are both operating effectively for the task at hand. Nevertheless, selecting the appropriate model requires careful consideration of many performance indicators as well as comprehension of the work context. For instance, you would favor a lower false positive rate model if the application has a high false positive cost. positive rate if the application has a high false positive cost. Discussion of results shows numerous classification model performance points:

The model excels in 'Email' and 'Resume,' with 90.0% and 96.7% true positive rates, respectively. This suggests that the model captures email and resume features and distinguishes them from other document types .Possible Class Confusion: Some classes are confusing. With 25.4% of notes misclassified as letters, 'Letter' and 'Note' are often confused. This could mean the features used to identify these classes are weak, or the document types are similar structurally and content-wise. Forms are sometimes misclassified as ADVE (5.6%) and Scientific materials as Reports (10.0%). These patterns can help diagnose feature selection and class imbalance issues. Class imbalance: Some classes have far more examples than others, possibly contributing to misclassification rates. This can bias the model toward the dominant classes. Strengths of the model: The diagonal of the matrix displays the model's strengths, with several classes predicted with over 80% accuracy. The model appears to have mastered most document types' traits.

Ahmed H. Salman, Waleed A, Mahmoud Al-Jawher. Image Document Classification Prediction based on SVM and gradient-boosting Algorithms. *Journal port Science Research*, 6(4), pp. 348-356. https://doi.org/10.36371/port.2023.4.5

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

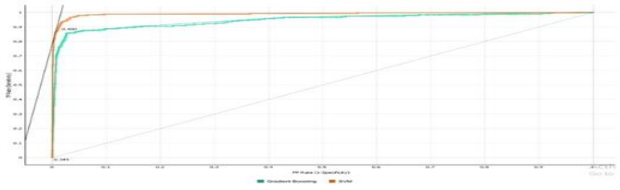

Fig 2 *Receiver Operating Characteristic (ROC) curve after using Support Vector Machines (SVM) and Gradient Boosting.*

Table 3: Compare The Suggested Approach To Related Works.

No Reference	Feature selection method	Classifier Method	Highest performance in accuracy
[1] (2015)	a deep Convolutional Neural Network (CNN)	adapt deep Convolutional Neural Networks (CNN)	Our suggested method has demonstrated superior performance compared to the current leading method on the Tobbacoo3428 dataset. It attained a median accuracy of 77.6% using 100 samples per class for training and validation.
[8] (2019)	deep CNN features for representing document images	CNN	The CNN, which was explicitly trained to classify only letters and memos, obtained a remarkable accuracy rate of 95% on this particular job.
[28] (2019)	CNN-based classification methods	k-means algorithm	90.04
[30] (2016)	CNN	lightweight Deep Convolutional Neural Network (DCNN) And SVM	The ensemble model achieved a median accuracy of 72.1% on the standard Tobacco3482 document dataset, based on 10 trials.
[31](2019)	Uses deep CNN models to extract structural features of document images.	InceptionV3 And the second stream, using an OCR	95.8
The proposed mothed (2023)	feature extraction or data embedding, we mean that we use the layers of the VGG-19 model to transform input images into a more compact and meaningful representation.	AdaBoost and SVM	SVM is achieving accuracy (0.964) while AdaBoost is achieving accuracy (0.853)

5. CONCLUSION AND FUTURE WORK

We presented an efficient multimodal mech learning multi-classifier for the classification of document images. We demonstrate that the models perform adequately even with limited data. Our studies involve training the suggested model using the Tobacco-3482 dataset. We achieved a level of accuracy of 0.964% for SVM and 0.853 for the AdaBoost Algorithm. In the future, we are planning To make the model better, one could think about acquiring extra training data for classes that aren't performing well. Improve the feature extraction procedure to capture the unique characteristics of the documents better. We are employing methods, such as resampling or class weights in model training, to address class imbalance. Investigating more intricate models or groups of models that could better represent the subtle differences between classes. We are analyzing the misclassified documents through error analysis to find out why the model doesn't work in specific cases. The data scientist can improve the model's performance across all document classes by iteratively resolving these issues.

As future propole a hybrid transform architecture [37] can be used that can do all the types of tasks [38] required and other sensors [39] that could make a much more General Artificial Intelligence model [40]. From our work: we specify a method for artificial general intelligence that can simulate human intelligence, implemented by taking in any form of arbitrary input data [41] the method comprising Learning to transform the arbitrary input data into an internal numerical format [42]. Then performing a plurality of numerical operations, the plurality of numerical operations comprises learned and

Ahmed H. Salman, Waleed A, Mahmoud Al-Jawher. Image Document Classification Prediction based on SVM and gradient-boosting Algorithms. *Journal port Science Research*, 6(4), pp. 348-356. https://doi.org/10.36371/port.2023.4.5

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

neural network operations, on the arbitrary input data in the internal format [43]. Then transforming the arbitrary input data into output data having output formats using a reciprocal process learned to transform the output data from the arbitrary input data, wherein all steps being done unsupervised[46-44]

ACKNOWLEDGEMENT

The authors express their gratitude to the 1Informatics Institute for Postgraduate Studies, Iraqi Commission for Computers and Informatics, located in Baghdad, Iraq, for their help in the current research.

REFERENCES

[1] Kadhim, A.I. Survey on supervised machine learning techniques for automatic text classification. Artif. Intell. Rev. 2019, 52, 273–292 .

[2] Kumbhar, P.; Mali, M.A. Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification. Int. J. Sci. Res. 2016, 5, 1267–1275 .

[3] Mwadulo, M.W. A Review on Feature Selection Methods for Classification Tasks. Int. J. Comput. Appl. Technol. Res. 2016, 5, 395–402 .

[4] Zhang, T.; Yang, B. Big data dimension reduction using PCA. In Proceedings of the 2016 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, USA, 18–20 November 2016; pp. 152–157 .

[5] Al-Anzi F. S., & AbuZeina D. (2017), 'Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing,' Journal of King Saud University – Computer and Information Sciences, 29(2), 189–195.

[6] I M Rabbimov, S Kobilov," Multiclass Text Classi_cation of Uzbek News Articles using Machine Learning," Journal of Physics: Conference Series 1546 (2020) 012097, IOP Publishing. doi:10.1088/1742-6596/1546/1/012097

[7] P. Dhar, M. Abedin, et al., Bengali news headline categorization using optimized machine learning pipeline, I. J. of Info. Eng. & Elec. Busi. 13 (1 .(2021) (

[8] M. M. H. Shahin, T. Ahmmed, S. H. Piyal, M. Shopon, Classification of Bangla news articles using bidirectional long short term memory, in IEEE TENSYMP, IEEE, 2020, pp. 1547–1551.

[9] Du C, Huang L. (2018) Text classification research with attention-based recurrent neural networks. International Journal of Computers Communications & Control, 13(1): 50-64. https://doi.org/10.15837/ijccc.2018.1.3142

[11] Muhittin IŞIK1, Hasan DAĞ," The impact of text preprocessing on the prediction of review ratings, Turk J Elec Eng & Comp Sci (2020) 28: 1405 – 1421. doi:10.3906/elk-1907-46.

[12] Mustafa Abdalrassual Jassim, Dhafar Hamed Abd and, Mohamed Nazih Omri," Machine Learning-based Opinion Extraction Approach from Movie Reviews for Sentiment Analysis, Research Square,(2022) DOI: https://doi.org/10.21203/rs.3.rs-1780497/v1

[13] Jwan k. Alwan, Abir Jaafar Hussain, Dhafar Hamed Abd, Ahmed Tariq Sadiq, Mohamed Khalaf, and Panos Liatsis," Political Arabic Articles Orientation Using Rough Set Theory With Sentiment Lexicon," IEEE Access (Vol 9),p.p24475 – 24484, DOI: 10.1109/ACCESS.2021.3054919.

[14] Adrita Barua, Omar Sharif, Mohammed Moshiul Hoque∗ ," Multiclass Sports News Categorization using Machine Learning Techniques: Resource Creation and Evaluation," 10th International Young Scientists Conference on Computational Science, Procedia Computer Science 193 (2021) 112–121. DOI: 10.1016/j.procs.2021.11.002.

[15] Abd, D.H., Sadiq, A.T., Abbas, A.R. (2020). Classifying Political Arabic Articles Using Support Vector Machine with Different Feature Extraction. In: Khalaf, M., Al-Jumeily, D., Lisitsa, A. (eds) Applied Computing to Support Industry: Innovation and Technology. ACRIT 2019. Communications in Computer and Information Science, vol 1174. Springer, Cham. https://doi.org/10.1007/978-3-030-38752-5_7 .

[16] Jannik Strtgen, Leon Derczynski, Ricardo Campos, Omar Alonso. Time and information retrieval: introduction to the special issue. Inf Process Manage 2015; 51(6): 786–90.

[17] Rajit Nair, Amit Bhagat, "Feature Selection Method To Improve The Accuracy of Classification Algorithm," International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-6, April 2019.

[18] Yechuang Wang, Penghong Wang, Jiangjiang Zhang, Zhihua Cui, Xingjuan Cai, Wensheng Zhang, and Jinjun Chen, "A Novel Bat Algorithm with Multiple Strategies Coupling for Numerical Optimization", Mathematics 2019, 7(2), 135; https://doi.org/10.3390/math7020135.

[19] TU Chengsheng, LIU Huacheng, XU Bing, "AdaBoost typical Algorithm and its application research", MATEC Web of Conferences 139, 00222 (2017). DOI: 10.1051/matecconf/201713900222.

[20] Hancock, J.T., Khoshgoftaar, T.M. CatBoost for big data: an interdisciplinary review. J Big Data 7, 94 (2020). https://doi.org/10.1186/s40537-020-00369-8 .

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

[21] Tian, Y.; Zhang, Y.; Zhang, H. Recent Advances in Stochastic Gradient Descent in Deep Learning. Mathematics 2023, 11, 682. https://doi.org/10.3390/math11030682.

[22] N. N. A. Sjarif, N. F. M. Azmi, S. Chuprat, H. M. Sarkan, Y. Yahya, S. M. Sam, SMS Spam message detection using term frequency-inverse document frequency and random forest algorithm, Proc. Comput. Sci., 161 (2019) 509-515. https://doi.org/10.1016/j.procs.2019.11.150 .

[23] M. Sarnovský, V. Maslej-Krešňáková, and N. Hrabovská, "Annotated dataset for the fake news classification in Slovak language," in 2020 18th International Conference on Emerging eLearning Technologies and Applications (ICETA), 2020, pp. 574-579: IEEE. DOI:10.1109/ICETA51985.2020.9379254.

[24] Al-Mejibli, I. S., Abd, D. H., Alwan, J. K., & Rabash, A. J. (2018). Performance Evaluation of Kernels in Support Vector Machine. 2018 1st Annual International Conference on Information and Sciences (AiCIS). doi:10.1109/aicis.2018.00029

[25] Khalaf, M., Hussain, A. J., Keight, R., Al-Jumeily, D., Keenan, R., Chalmers, C., … Idowu, I. O. (2017). Recurrent Neural Network Architectures for Analysing Biomedical Data Sets. 2017 10th International Conference on Developments in eSystems Engineering (DeSE). doi:10.1109/dese.2017.12.

[26] Abd, Dhafar Hamed; Alwan, Jwan K.; Ibrahim, Mohamed; Naeem, Mohammad B. (2017), The utilization of machine learning approaches for medical data classification and personal care system management for sickle cell disease, IEEE 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT), 213–218. doi:10.1109/ntict.2017.7976147

[27] Ibraheem Al-Jadir, Waleed A Mahmoud "A Grey Wolf Optimizer Feature Selection method and its Effect on the Performance of Document Classification Problem" journal port science research, Vol. 4, Issue 2, PP. 125-131, 2021

[28] Waleed A Mahmoud, MR Shaker "3D Ear Print Authentication using 3D Radon Transform" proceeding of 2nd International Conference on Information & Communication Technologies, Pages 1052-1056, 2006.

[29] RA Dihin, E AlShemmary, W Al-Jawher "Diabetic Retinopathy Classification Using Swin Transformer with Multi Wavelet" Journal of Kufa for Mathematics and Computer 10 (2), 167-172, 2023.

[30] Rasha. A. Dihin,2Ebtesam N. AlShemmary, 3Waleed Ameen Al Jawher "Implementation Of The Swin Transformer and Its Application In Image Classification" Journal Port Science Research 6 (4), 318-331.

[31] Ali Akram Abdul-Kareem, Waleed Ameen Mahmoud Al-Jawher, "Image Encryption Algorithm Based on Arnold Transform and Chaos Theory in the Multi-wavelet Domain", International Journal of Computers and Applications, Vol. 45, Issue 4, pp. 306-322, 2023.

[32] Sarah H Awad Waleed A Mahmoud Al-Jawher, "Precise Classification of Brain Magnetic Resonance Imaging (MRIs) using Gray Wolf Optimization (GWO)" HSOA Journal of Brain & Neuroscience Research. Vol 6, issue 1, pp 100021, 2022

[33] Ali Akram Abdul-Kareem, W. A. Mahmoud Al-Jawher, Zahraa A Hasan, Suha M. Hadi," Time Domain Speech Scrambler Based on Particle Swarm Optimization" International Journal for Engineering and Information Sciences, Vol. 18, Issue 1, PP. 161-166, 2023.

[34] Zahraa A Hasan, Suha M Hadi, Waleed A Mahmoud, "Speech scrambler with multiwavelet, Arnold Transform and particle swarm optimization" Journal Pollack Periodica, Volume 18, Issue 3, Pages 125-131, 2023.

[35] Ali Akram Abdul-Kareem, Waleed Ameen Mahmoud Al-Jawher, "WAM 3D discrete chaotic map for secure communication applications" International Journal of Innovative Computing, Volume 13, Issue 1-2, Pages 45-54, 2022.

[35] Ali Akram Abdul-Kareem, Waleed Ameen Mahmoud Al-Jawher" Hybrid image encryption algorithm based on compressive sensing, gray wolf optimization, and chaos" Journal of Electronic Imaging, Volume 32, Issue 4, Pages 043038-043038, 2023.

[37] Waleed A. Mahmud Al-Jawhar "Feature Combination & Mapping Using Multiwavelet Transform" Al-Rafidain Univ. College journal, Issue 19, Pages: 13-34, 2006.

[38] Waleed Ameen Mahmoud "A Smart Single Matrix Realization of Fast Walidlet Transform" Journal International Journal of Research and Reviews in Computer Science, Volume 2, Issue, 1, Pages 144-151, 2011.

[39] Hamid M Hasan, Waleed A. Mahmoud Al- Jawher, Majid A Alwan "3-d face recognition using improved 3d mixed transform" Journal International Journal of Biometrics and Bioinformatics (IJBB), Volume 6, Issue 1, Pages 278-290, 2012.

[40] AHM Al-Heladi, W. A. Mahmoud, HA Hali, AF Fadhel "Multispectral Image Fusion using Walidlet Transform" Advances in Modelling and Analysis B, vol 52, issue 1-2, pp. 1-20, 2009.

[41] Walid A Mahmoud, Majed E Alneby, Wael H Zayer "2D-multiwavelet transform 2D-two activation function wavelet network-based face recognition" J. Appl. Sci. Res, vol. 6, issue 8, 1019-1028, 2010.

[42] Waleed A Mahmoud, MR Shaker "3D Ear Print Authentication using 3D Radon Transform" proceeding of 2nd International Conference on Information & Communication Technologies, Pages 1052-1056, 2006.

[43] Waleed A Mahmoud, Afrah Loay Mohammed Rasheed "3D Image Denoising by Using 3D Multiwavelet" AL-Mustansiriya J. Sci, vol 21, issue 7, pp. 108-136, 2010.

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 6, No:4. 2023

[44] Maryam I Mousa Al-Khuzaay, Waleed A Mahmoud Al-Jawher, "New Proposed Mixed Transforms: CAW and FAW and Their Application in Medical Image Classification" International Journal of Innovative Computing, Volume 13, Issue 1-2, Pages 15-21, 2022.

[45] Hamid M Hasan, Waleed Ameen Mahmoud Al-Jawher and Majid A Alwan, "3-d face recognition using improved 3d mixed transform" International Journal of Biometrics and Bioinformatics (IJBB), Volume 6, Issue 1, Pages 278, 2012.

[46] WA Mahmoud, RA Jassim, " Image Denoising Using Hybrid Transforms" Engineering and Technology Journal, Volume 25, Issue 5, Pages 669-682, 2007.

Ahmed H. Salman, Waleed A, Mahmoud Al-Jawher. Image Document Classification Prediction based on SVM and gradient-boosting Algorithms. *Journal port Science Research*, 6(4), pp. 348-356. https://doi.org/10.36371/port.2023.4.5