BANK DOCUMENT CLASSIFICATION

A Project Report Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

P BUJJI	(N190422)
B SRUTHI	(N190897)
B NAGARANI	(N190201)
B PRAVALLIKA	(N190396)
R UMADEVI	(N191106)
G SIDDARDHA	(N190392)

Under the Esteem Guidance of

Mrs. M. J. Blessy

Assistant Professor,CSE Dept

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Rajiv Gandhi University of Knowledge Technologies – Nuzvid

Nuzvid, Eluru, Andhra Pradesh – 521202.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES

(A.P.Government Act 18 of 2008)

RGUKT-NUZVID, Eluru Dist.-521202

Tele Fax:08656-235557/235150

CERTIFICATE OF COMPLETION

This is to certify that the work entitled, **“Bank Document Classification”** is the bonafide work of **B. PRAVALLIKA (ID No: N190396)**, **B. SRUTHI (ID No: N190897)**, **B. NAGARANI (ID No: N190201)**, **P. BUJJI (ID No: N190422)**,**R. UMADEVI (ID No: N191106)** and **G. SIDDARDHA (ID No: N190392)** carried out under my guidance and supervision for the 4th year project of **Bachelor of Technology** in the Department of Computer Science and Engineering under RGUKT Nuzvid. This work is done during the academic session **July 2024 – April 2025**, under our guidance.

. . . .

Mrs. M. J. Blessy
Assistant Professor,
Department of CSE
RGUKT-Nuzvid

. . . .

Mrs. S. Bhavani
Assistant Professor,
Head of the Department-CSE,
RGUKT-Nuzvid

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES

(A.P.Government Act 18 of 2008)

RGUKT-NUZVID,Eluru Dist.-521202

Tele Fax:08656-235557/235150

CERTIFICATE OF EXAMINATION

This is to certify that the work entitled, **"Bank Document Classification"** is the bonafide work of **B. PRAVALLIKA (ID No: N190396)**, **B. SRUTHI (ID No: N190897)**, **B. NAGARANI (ID No: N190201)**, **P. BUJJI (ID No: N190422)**,**R. UMADEVI (ID No: N191106)** and **G. SIDDARDHA (ID No: N190392)**. We hereby accord our approval of it as a study carried out and presented in a manner required for its acceptance in the 4th year of **Bachelor of Technology** for which it has been submitted.

This approval does not necessarily endorse or accept every statement made, every opinion expressed or conclusion drawn, as recorded in this thesis. It only signifies the acceptance of this thesis for the purpose for which it has been submitted.

....	....
Mrs. M. J. Blessy	**Project Examiner**
Assistant Professor,	Department of CSE,
Department of CSE,	RGUKT-Nuzvid
RGUKT-Nuzvid	

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES

(A.P.Government Act 18 of 2008)

RGUKT-NUZVID,Eluru Dist.-521202

Tele Fax:08656-235557/235150

DECLARATION

We, **B. Pravallika** (ID No : N190396), **B. Sruthi** (ID No : N190897), **B. Nagarani** (ID No : N190201), **P. Bujji** (ID No : N190422), **R. Umadevi** (ID No : N191106), and **G. Siddardha** (ID No: N190392), hereby declare that the project report entitled **"Bank Document Classification"**, done by us under the guidance of **Mrs. M. J. Blessy**, Assistant Professor, is submitted for the partial fulfillment for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** during the academic session **July 2024 – April 2025**, at **RGUKT-Nuzvid**.

We also declare that this project is a result of our own effort and has not been copied or imitated from any source. Citations from any websites are mentioned in the references. The results embodied in this project report have not been submitted to any other university or institute for the award of any degree or diploma.

Date: 26-04-2025
Place: Nuzvid

P. BUJJI	**(N190422)**
B. SRUTHI	**(N190897)**
B. NAGARANI	**(N190201)**
B. PRAVALLIKA	**(N190396)**
R. UMADEVI	**(N191106)**
G. SIDDARDHA	**(N190392)**

ACKNOWLEDGEMENT

We would like to express our profound gratitude and deep regards to our respected supportive guide **Mrs.M.J.Blessy** for her exemplary guidance, monitoring and constant encouragement to us throughout the B.Tech course. We shall always cherish the time spent with her during the course of this work due to the invaluable knowledge gained in the field of reliability engineering.

We are extremely grateful for the confidence bestowed in us and entrusting our project entitled **“Bank Document Classification”**.

We express gratitude to **Mrs. S. Bhavani** (HOD of CSE) and other faculty members for being a source of inspiration and constant encouragement which helped us in completing the project successfully.

Finally, yet importantly, we would like to express our heartfelt thanks to our beloved God and parents for their blessings, our friends for their help and wishes for the successful completion of this project.

ABSTRACT

Bank Document Classification using Ensemble Machine Learning Techniques

Banks face challenges in classifying the diverse array of financial documents they handle daily, making manual classification time-consuming, error-prone and resource-intensive. The inconsistency in document formats and the need for accurate data extraction complicate the process, leading to inefficiencies and increased operational costs. ML techniques used for bank document classification can improve the accuracy and speed of document categorization, helping organizations save time, resources, and money by automating the process and reducing manual effort. Applications include Automated Document Management, Fraud Detection, Customer Service, and Loan Processing. Existing solutions typically use classical algorithms, which may not fully leverage the combination of text and image analysis, resulting in lower accuracy, slower processing times, and increased manual intervention. The project uses modern machine learning algorithms such as XGBoost, CatBoost, and Voting Classifier for classification. It combines text and image analysis through Computer Vision, Natural Language Processing, and Deep Learning techniques to achieve higher accuracy and efficiency. The approach involves extracting text from images and processing it before feeding it into the algorithms. The dataset consists of various financial documents, including bank statements, invoices, credit card statements, and tax returns. It contains both text and image data, where the text is extracted from images for comprehensive analysis. The integration of XGBoost, CatBoost, and Voting Classifier algorithms, along with Computer Vision and Natural Language Processing, allows for faster and more accurate document classification. This significantly reduces manual effort and improves the efficiency of the document categorization process for banks.

Table of Contents

List of figures

List of Tables

1. INTRODUCTION

1.1. What is the problem?

Banks face significant challenges in classifying the vast array of financial documents they handle daily, such as bank statements, invoices, credit card statements, and tax returns. The sheer volume and diversity of these documents make manual classification an arduous task. Each document type often has a different format, which adds complexity to the classification process. Manual classification is not only time-consuming but also prone to human errors, leading to inaccuracies. These inaccuracies can have serious repercussions, including financial losses and compliance issues. The manual process also requires substantial human resources, increasing operational costs. Additionally, inconsistencies in document formats further complicate data extraction and processing. The need for accurate data extraction is crucial, as errors can lead to incorrect classifications and subsequent issues. Overall, these challenges result in inefficiency and high operational costs for banks.Therefore, there is a pressing need for an automated, efficient, and accurate document classification system.

1.2. Its applications

Machine learning (ML) techniques for bank document classification enhance operational efficiency through several applications. Automated document management reduces manual sorting and minimizes human error, while fraud detection identifies irregularities by analyzing document patterns. In customer service, ML-driven classification enables quicker and more accurate responses to financial document queries. Loan processing benefits from streamlined document verification approval and compliance is supported to ensure correct categorization and storage according to regulatory requirements, reducing the risk of noncompliance. These improvements in speed and accuracy enhance overall decision-making, allowing banks to focus on strategic initiatives, ultimately leading to enhanced customer satisfaction and operational performance.

1.3. Limitations of the existing solutions

Current document classification solutions in banks often rely on classical algorithms, which have several limitations. These classical algorithms typically analyze text or image data separately, failing to leverage the full potential of combining both. As a result, the accuracy of classification is often compromised. Classical algorithms may also struggle with the diversity and complexity of financial documents, leading to slower processing times. This inefficiency can necessitate increased manual intervention, negate the benefits of automation. Additionally, these algorithms might not adapt well to variations in document formats, causing inconsistencies in classification results. They also lack the advanced capabilities required to handle large-scale document datasets effectively. As a result, the operational costs and resource requirements remain high.Furthermore, the inability to process documents swiftly and accurately can lead to delays and errors in banking operations. These limitations highlight the need for more advanced and robust machine learning techniques. Consequently, banks must seek innovative solutions to overcome these challenges and improve document classification efficiency.

1.4. What is your approach and based on what?

Our approach leverages modern machine learning algorithms, including XGBoost, CatBoost, and Voting Classifier, to enhance document classification in banks. These algorithms are chosen for their robustness, accuracy, and efficiency in handling complex datasets. We combine text and image analysis through advanced techniques like Computer Vision and Natural Language Processing (NLP). This holistic approach allows us to extract meaningful insights from both textual and visual data in financial documents. Computer Vision is utilized to process and interpret image data, identifying key features and patterns. NLP techniques are employed to analyze text data, extracting relevant information and keywords. By integrating Deep Learning, we further enhance the accuracy of our classification model. This approach not only improves classification precision but also speeds up the processing time. The combination of these advanced techniques addresses the limitations of classical algorithms.Our system is designed to adapt to various document formats, ensuring consistent and accurate classification. Overall, our approach provides a comprehensive solution for efficient and accurate document classification.

1.5. On which dataset?

The dataset for this project comprises a diverse collection of financial documents, crucial for training and testing our classification model. It includes bank statements, invoices, credit card statements, and tax returns, representing a broad spectrum of document types banks handle. This dataset is designed to contain both text and image data, reflecting real-world scenarios where documents come in various formats. Text data is extracted from images using Optical Character Recognition (OCR) techniques, ensuring that the model can analyze both textual and visual information. The inclusion of diverse document types helps the model learn to identify and classify different categories accurately. Each document in the dataset is labeled, providing a clear indication of its type, which is essential for supervised learning. This labeling enables the model to learn the distinguishing features of each document.

The dataset is also preprocessed to remove noise and standardize the formats, enhancing the quality of the training data. By using such a comprehensive dataset, we ensure that the model is robust and performs well on real-world documents. The diversity and quality of the dataset are critical for achieving high accuracy and reliability in document classification.

1.6. Result

The integration of advanced algorithms like XGBoost, CatBoost, and Voting Classifier, combined with Computer Vision and Natural Language Processing, has yielded impressive results in document classification. Our system achieves faster and more accurate classification of financial documents, significantly reducing the need for manual effort. The use of modern machine learning techniques allows for better handling of diverse and complex document formats. By extracting and processing text from images, our approach ensures comprehensive analysis and accurate categorization. This leads to increased operational efficiency and reduced processing times for banks. The accuracy of classification is notably higher compared to classical algorithms, minimizing errors and enhancing decision-making. Additionally, the automated system supports various banking operations, including fraud detection, customer service, and loan processing. Overall, our approach not only improves classification efficiency but also contributes cost savings and improved service quality.The successful implementation of this system demonstrates the potential of modern ML techniques in transforming bank document management.

2. RELATED / EXISTING WORK

Sarosh Dandoti (2022) proposes the process of categorizing each document based on the information. Designing a multi-label classification model with parameter tuning to enhance performance and forecasts is the aim of this study. The author used multiple Machine Learning models such as SVMs, Naïve Bayes, Random Forest, and KNN. The method of categorizing documents according to their content is known as document classification. Therefore, it is more cost-effective and accurate to perform document classification using machine learning models.

The structure of the document can be recognized by ML models using OCR, which allows for the classification of documents into various sections. The author tuned the hyperparameters of each ML model, including KNNs, Naive Bayes, SVMs, and Random Forest. The models have been tuned using GridSearchCV and KFold Cross-Validation. The proposed model can be used to determine whether the provided text contains hate speech and unnecessary material.

Arslan, Ö., & Uymaz, S. A. (2022) suggested a CNN-based approach for classifying invoice photos as financial documents. Four separate classes of invoice photographs were utilised to construct the data set for the system, and the images were collected from a Turkish bank's database. The Hough transform is used to find lines in images, but it may also be used to find the locations of arbitrary forms. The image is resized using the zero-padding technique by adding fixed-value pixels to the image's edges.

To prevent the issue of lost detail when scaling photographs in various aspect ratios, the BAFGV2 dataset was created. The Original data set's photos were converted into the BAFGV2 data set using the zero-padding technique. Based on the original data, CNN architecture had the highest accuracy rates. Using three different sets of data, the LeNet-5, VGG-19, and MobileNetV2 architectures were trained. The BAFGV3 dataset, which was created by extending the BAFGV2 dataset using varying brightness ratios, has the greatest accuracy rates. The VGG-19 architecture used in the BAFGV3 dataset resulted in the highest success rate, which is the greatest test success rate recorded across all experiments.

Sheth, V. et al. (2022) provides a thorough review of the literature on several machine learning methods used to solve classification problems. The first section of the paper discusses the significance of categorization tasks in machine learning and the difficulties involved. The authors experimented with different machine learning methods, including Decision Trees, Naive Bayes, Logistic Regression, K-Nearest Neighbors, Support Vector

Machines, and Neural Networks.

Later, they compared these algorithms based on various performance indicators, including precision, accuracy, recall, F1-score, and AUC-ROC.

They also discussed how the performance of these algorithms is impacted by a number of variables, including dataset size, feature choice, and hyperparameter adjustment. Finally, the paper gives a thorough analysis of numerous machine learning methods utilized for categorization problems. Naive Bayes classifiers, a type of classification technique, are produced using Bayes' theorem. It is discovered that, in terms of accuracy, recall, precision, and F1 score, the Naive Bayesian classification model outperforms the competition.

Trivedi, S. et al. (2020) discussed a document classification scheme, where the Naive Bayes technique, a machine learning algorithm, is used to perform this categorization. It has a number of phases, and each phase can be completed using several methods. This study classifies texts using machine learning techniques. Systems can 6 recognise patterns using machine learning by using current algorithms and data sets are trained using a variety of techniques in machine learning. The Naive Bayes algorithm is utilised in this study because, when compared to other algorithms, it exhibits the maximum efficiency, is simple to construct, and is especially helpful for very big data sets.

Techniques for classifying texts that are often employed include support vector machines, knearest neighbours, and the nave Bayesian method. Naive Bayes, KNN, and SVM perform best with small sample sizes, whereas Decision Tree performs best with larger samples. This work highlights the effectiveness of using Naive Bayes for document classification. One of the machine learning algorithms, the Naive-Bayes technique, will be used to conduct classification. It accepts input in the form of PDF documents that are categorised according to their domain. Employing machine learning, only PDF-formatted documents are supported.

Ghumade, T et al.,(2019) discussed the value of document classification in several contexts, including spam filtering, sentiment analysis, and topic modelling. The authors go on to give a general review of several methods used for document classification, including rule-based approaches, machine learning algorithms, and deep learning models. They give an overview of NLP methods like lemmatization, stemming, stop-word elimination, and tokenization. The study also provides a review of numerous studies that have classified documents using NLP and RNNs. This study offers a thorough assessment of the literature on the various methods for document classification and concentrates on the application of NLP and RNNs to this task. It can be a helpful tool for academics and professionals involved in document classification, machine learning, and natural language processing.

Engin, D. et al. (2019), proposed a deep neural network for classifying customerordered petition-based publications.These texts' structures vary in terms of their use of tables, lines, and various formats. The data is extracted from the documents using optical character recognition (OCR) technology.

The extraction of text from photos and multimodal documents has been the focus of deep learning-based methods.A specific area of a document is employed as a feature extractor by a Convolutional Neural Network (CNN) .The multimodal categorization proposed Deep Neural Networks on both text and image formats. The CNN architecture for image categories to retrieve information from images and which was before (LSTM) model to do same for text inputs.CNN architectures have already been suggested and data augmentation has been used. CNN architectures have already been suggested and data augmentation has been used. AlexNet A. Krizhevsky(2012), VGG-16 K. Simonyan(2014), GoogLeNet, . Szegedy et al(2015) and Resnet K. He, X. Zhang(2016) are a few well-known CNN architectures that have been examined for document classification utilising transfer learning in M. Z. Afzal, (2017). Learning visual and linguistic elements allows for the classification of all petition-based client order documents. The proposed system's goal is to categorise both pictures and text documents. using a pre-trained LSTM model to obtain data from text data and a well before CNN model for sight modality to collect data from pictures. Several sensations, including late fusion and early fusion, were used in multimodal categorization. Whereas late fusion approaches are decision-driven, preliminary fusion techniques depend on extracting features for several modalities. The suggested system used two approaches to classify multimodal documents. Learning visual and linguistic elements allows for the classification of all petition-based client order documents.

Sharma, K et al.(2018) proposed deep learning techniques to expedite lassification and suggest pertinent documents. Recurrent Neural Network with Convolutional Neural 8 Network, a suggested deep learning technique, aids in building a reliable classifier model using a range of training data. The proposed system uses deep learning methods to speed up classification and provide useful document suggestions. A recommended deep learning method, recurrent neural network with convolution neural network, helps create a trustworthy classifier model employing a variety of training data. A convolution neural network (CNN) and recurrent neural network (RNN) based method is capable of efficiently expressing textual features and modelling highorder label correlation with a reasonable computational complexity, according to Guibin Chen, Deheng Ye, and Zhenchang Xing(2017). A approach based on TF-IDF, a numerical statistic meant to reflect how essential a word is to a document in a collection or corpus, is suggested by Md. Saiful Islam (2017).This is accomplished by combining natural language processing with deep learning models (CNN, RNN, RCNN) and SVM (Support Vector Machine) approaches. The variety of documents in the dataset can be enhanced to improve classifier accuracy. To ensure that semantic correlations are taken into account during classification, the Convolutional and Recurrent Neural Network (RCNN) algorithm is used. Cüceloğlu, İ., Oğul, H introduced a system for classifying papers that are regularly submitted with bank applications.

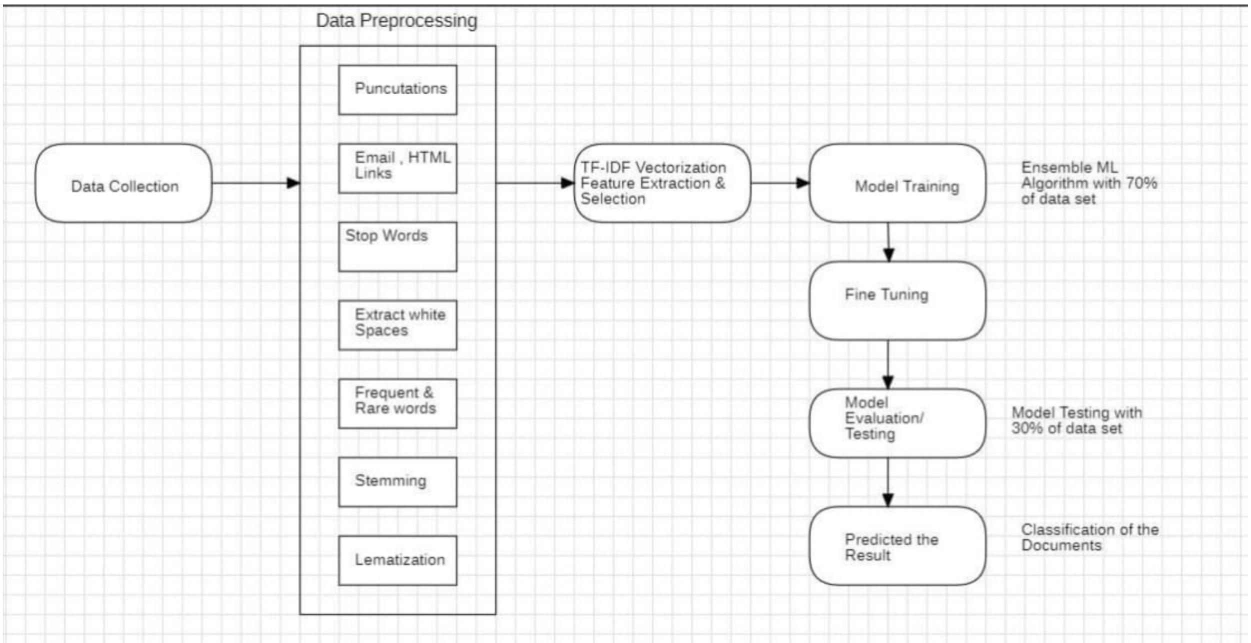
The retrieved text information serves as only foundation for the framework. Turkish-specific feature extraction and selection methods are used. We create a system for classifying frequently used bank papers in Turkish banks in an attempt to deal with the second difficulty by just using the text content to represent images. For Turkish texts, an optical character recognition (OCR) machine is used after noise removal and page segmentation. Support Vector Machine (SVM) and Multinomial Bayesian Network are two widely used supervised learning methods that are taken into consideration for categorization (MNB). Basarkar(2017) explored that document classification is a persistent issue in information retrieval three categories can be used to broadly classify automatic document classification. These three types of categorization are semi-supervised,supervised and supervised document classification. The two key criteria that make document classification difficult are as follows: Topic ambiguity and (a) feature extraction.Among the methods used to categorise documents include Decision Trees, Support Vector Machines, Naive Bayes, Expedition maximisation, and Neural Networks. For the objective of extracting language features for a given topic, the features extraction algorithm combined two very independent and potentially antagonistic metrics: Popularity (a); rarity (b).To determine the rarity of any given phrase, they used the Lexical Data Consortium (LDC) data collection to study the probability of recurrence of any n-gram upon this Web. The authors used regression, Naive Bayes, Multi - layer perceptron, RBF networks, and other classification techniques in addition to random forest. A Multinomial Nave Bayes Classification model is created for each type of description (Binary Vectorizer, Count Vectorizer, and TfIdf Vectorizer).

Yaram, S.(2016) implemented Decision Tree, Random Forest, and Naive Bayes categorization algorithms, as well as applicable industry use cases, are the main topics of this study. Performance metrics like accuracy,precision and recall will be used to compare the 13 effectiveness of three classification methods. In machine learning, learning occurs from the input dataset, and predictions are made using the data provided 10 to the algorithm.

While using document clustering, a programme is given a large collection of data objects and is tasked with identifying patterns and links among them Nicholas O. Classification uses a supervised learning method Igor Kononenko and Matjaz Kukar, Yiming Yang and Jan O. Pedersen, where the relationship for a new dataset will be determined according to the learning already done. A dataset that is randomly split into two bundles and used as the testing set and testing dataset feeds all classification algorithms. The Hadoop MapReduce framework is used to handle the unstructured data after it has been placed into the Hadoop Distributed File System (HDFS).Feature extraction in supervised. An effort has been made in this work to highlight the key characteristics of machine learning algorithms, including Document classification and clustering (Decision Tree, Random Forest, and Nave Bayes) methods.

3. Proposed Methodology

3.1 Flowchart


Data Preprocessing
Puncutations
Email , HTML Links
Stop Words
Extract white Spaces
Frequent & Rare words
Stemming
Lematization
Data Collection
TF-IDF Vectorization Feature Extraction & Selection
Model Training
Ensemble ML Algorithm with 70% of data set
Fine Tuning
Model Evaluation/ Testing
Model Testing with 30% of data set
Predicted the Result
Classification of the Documents

3.1.a Architecture Diagram for Document Classification

3.2 Components

3.2.1 Data Description

Dataset contains set of bank documents which are categorized into different classes. Bank document classification involves categorizing different types of documents commonly used in banking operations into specific groups based on their content and purpose. Data is collected from various resources business stakeholders, customers, employees who submitted the tax returns collected from websites. The collected data is categories or classified and labeled into 4 types such as bank statements, invoices, credit card statements, and tax returns. The dataset consists of large collection of documents in different formats such as Images and PDF. These documents may be structured or unstructured they may contain lines, numbers, symbols, graphs, tables and images, The data for bank document classification typically consists of a large collection.

The dataset consists of documents in various formats such as PDF, Microsoft Word, and plain text. These documents may be structured or unstructured, meaning they may contain tables, graphs, or images in addition to text. The collected data is preprocessed by converting them into a uniform format. After preprocessing, the data is labeled with various categories based on the type of document to train the model for bank document classification. The model requires appropriate training data to classify the documents. The training data includes a variety of document types and formats, as well as accounting for any regional or language differences that may affect the classification process. The dataset preparation involves adding a variety of documents, removing irrelevant data, and modifying the data labels while the data can be refined for the model to classify the documents.

3.2.2. Preprocessing

Pre-processing is an essential step in classifying bank documents because it cleans and normalize data, making it easier for the model to classify documents accurately.Preprocessing of document classification involves the following steps for efficient classification. Data Collection: Acquire the relevant documents from various sources like business stakeholders and websites. Explore and Load the data by import all necessary packages or libraries for the preprocessing of the dataset. The documents convert into standard format by using some document conversation tools such as Optical Character Recognition (OCR) , Pypdf2 Cleaning the Data: Documents contain both relevant and irrelevant information. It is required to clean the data by removing irrelevant information such as stop words, punctuations, and special characters, headers, footers, extra spaces, html links and urls, email ids from the documents. This irrelevant information is not used for data analysis. Stop words are the common words in text such as ”a”, ”an”, ”the”, and ”and”, which do not contribute to the meaning of the text and interfere with the classification process.Remove the stop words using python library called NLTK, SpaCy.

Tokenization: Tokenization is the process of separating the text into words or sentences using the `NLTK` library. These tokens are useful for understanding context or developing NLP models.

Stemming: Stemming is a process which reduces a word to its base or root form. It is used to normalize the text.

Lemmatization: Lemmatization is a method of normalizing text documents. The main purpose of text normalization is to keep the vocabulary small and to remove noise, which helps improve the accuracy of many language modeling tasks.

3.2.3. Feature Selection and Extraction

Feature selection is a important step in data preprocessing for any machine learning task, including document classification. This involves selecting the most relevant subset of features from the raw data to use in the modeling process. This reduces the dimensionality of the data, improves model accuracy, and reduces computation time. The process of extracting new, more specific characteristics from raw data in order to retrieve the majority of its valuable information is known as feature extraction. Feature extraction involves selecting relevant features from the documents that are essential for classification.

Vectorization: In general, Machines can't understand or processed text data in a raw form. The text is converted into numerical format (Vector) that easily readable by the Machine. TF- IDF(Term frequency — Inverse document frequency) The effective technique for identifying the most crucial words in your text is the bag-ofwords model called TF-IDF (Term frequency — Inverse document frequency). Understanding the term frequency (TF) and inverse document frequency can help you understand the TF-IDF idea (IDF). The TF-IDF representation is a potent method for representing text data since it is the result of the combination of the TF and IDF representations.

Term Frequency (TF): As the name suggests, the term frequency gives the count of the terms present in a document with respect to a BoW. Term frequency is a function of term t and document d. TF(t,d) = (Number of occurrences of the term t in document d) (Number of terms present in the document d)

Inverse Document Frequency (IDF): A term's unique relationship to a document in relation to a corpus of documents is measured by inverse document frequency (IDF). The underlying assumption is that a phrase that is present in the majority of target documents in the corpus does not offer unique information. For each term in your Bag of Words, an explanation of inverse document frequency is provided. IDF(t) = log (Number of documents) (Number of documents that have term t) TfIdf (t,d,D) = Tf(t,d) * Idf(t,D) The data split into training and testing in the ratio of :0.70 0.30. N-Grams a continuous series of N items taken from a particular text or speech sample. A character, word, or sentence can be used as an item in this case, and N can be any number. The sequence is known as a bigram when N is 2. A trigram is the same as a sequence of three items, and so on. In this part, an N-Gram model at the word level is developed. 24 i. Save potential words in the n-gram dictionary and their frequency in the frequency dictionary so that they can later be sorted according to their frequency when sorting the dictionary values. ii. Iterate over each word, joining the n words that are already present to produce a N-gram.

3.2.4. Model Selection (Algorithms)

A classification algorithm is a machine learning algorithm used to predict a categorical or discrete target variable based on input features. These algorithms are used in various applications such as classification of documents, spam detection, sentiment analysis, image recognition, etc. Some of the popular machine learning algorithms for classifying documents are listed below

3.2.4.1. Classifier -1: Random Forest

Random Forest is one of the supervised machine learning algorithm can used for classification and regression. Random Forest is a machine learning algorithm that is widely used for classification, and regression. Several decision trees are used to construct the Random Forest method, and the decision is ultimately reached by the decision trees' majority votes. IIt is an approach to ensemble learning that integrates a number of decision trees to increase the model's efficacy. Each decision contains decision leaf nodes, decision node and root node. The leaf node of each decision tree is the final output. The final out of the Random Forest classifier chosen based on the majority of decision trees. iii.Choose samples at random from the dataset. iv. Create several decision trees using selected samples. v. Use each decision tree to make a prediction. vi. Use a voting system to choose the winning forecast.

3.2.4.2. Classifier-2: XGBoost Classifier

The ensemble classifier XGBoost applies the idea of gradient boosting. Boosting combines the performance of a number of weak learners to make weak learners stronger. In boosting, trees are formed sequentially, and each one lowers the inaccuracy of the ones before it. For the division of already-existing nodes, the XGBoost algorithm employs a greedy strategy. By including a regularisation term, it decreases the loss function.

3.2.4.3. Classifier-3: CatBoost Classifier

The CatBoost, or categorical boosting, is a potent method for regression and classification issues in machine learning that is based on gradient descent. CatBoost can be used for recommendation systems, ranking, and prediction. It can handle categorical data directly without extensive pre- processing.

3.2.4.4. Ensemble Model (Voting Classifier)

Voting classifier is an ensemble learning method that combines multiple models to make predictions. The purpose of voting classifier is to leverage the strengths of multiple models, each model has its own strengths and weaknesses to create more accurate prediction. The advantage of Voting Classifier is to produce higher accuracy than the individual model. The ensemble model reduce the errors made by individual model.

4. RESULTS AND DISCUSSION

4.1. Evaluation Parameters

The performance of the proposed system can be measured by several evaluation metrics.

Confusion Matrix:It represents classification results per various classification algorithms.

True Positive (TP): The number of positive occurrences that are considered as positive.

True Negative (TN):The number of negative occurrences that are considered as negative.

False Positive (FP):The number of negative occurrences that are considered as positive.

False Negative (FN):The number of positive occurrences that are considered as negative.

Accuracy:The classification performance for all classes is denoted by the following formula.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision:It talks about how accurate the model is in predicting false positive occurrences.

$$Precision = \frac{TP}{TP + FP}$$

Recall:Completeness of classifiers can be measured using Recall, which indicates how many of the actual positive instances are captured by the model as positive.

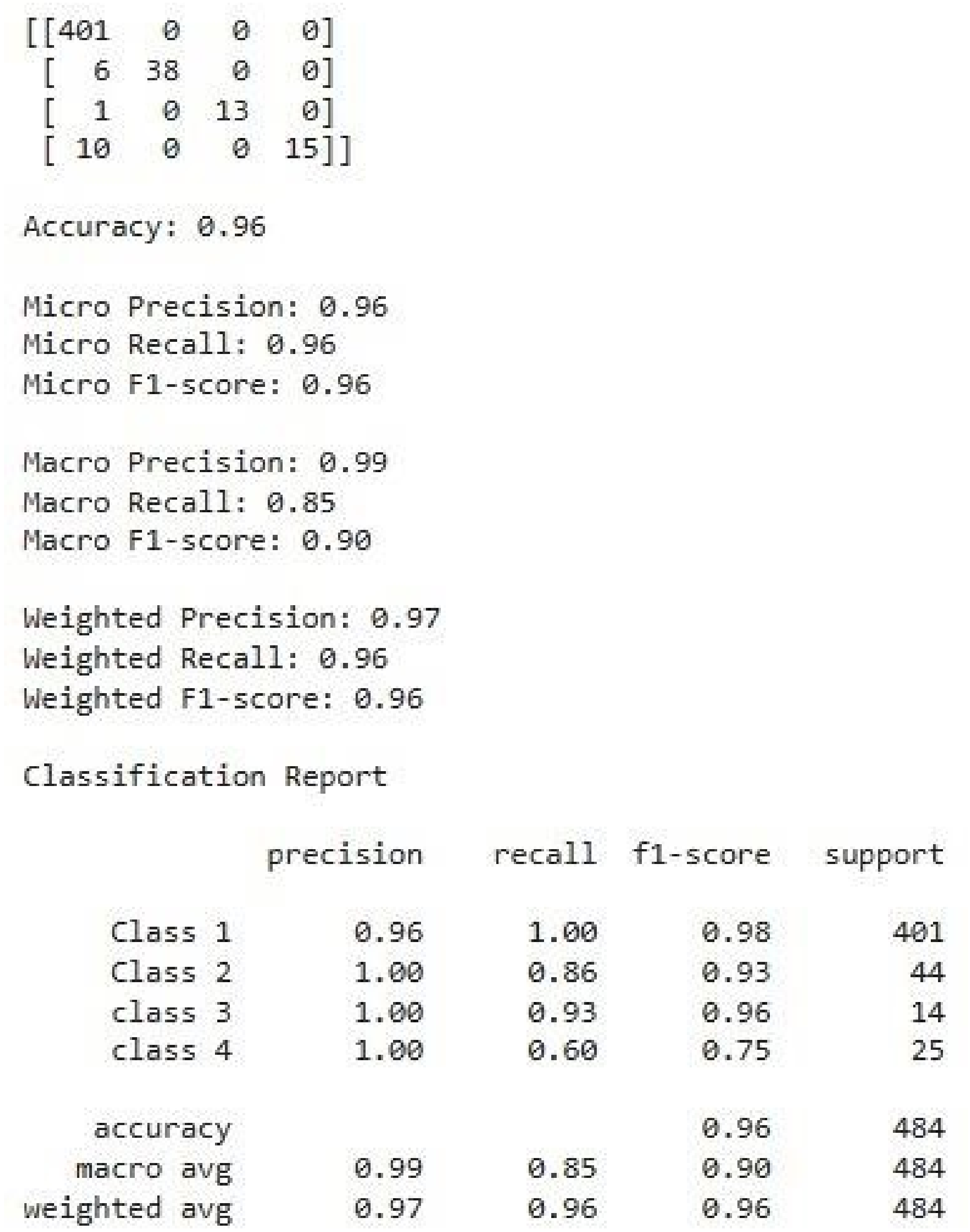
$$Recall = \frac{TP}{TP + FN}$$

F1-Score:The F1-score computes model accuracy that combines precision and recall.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

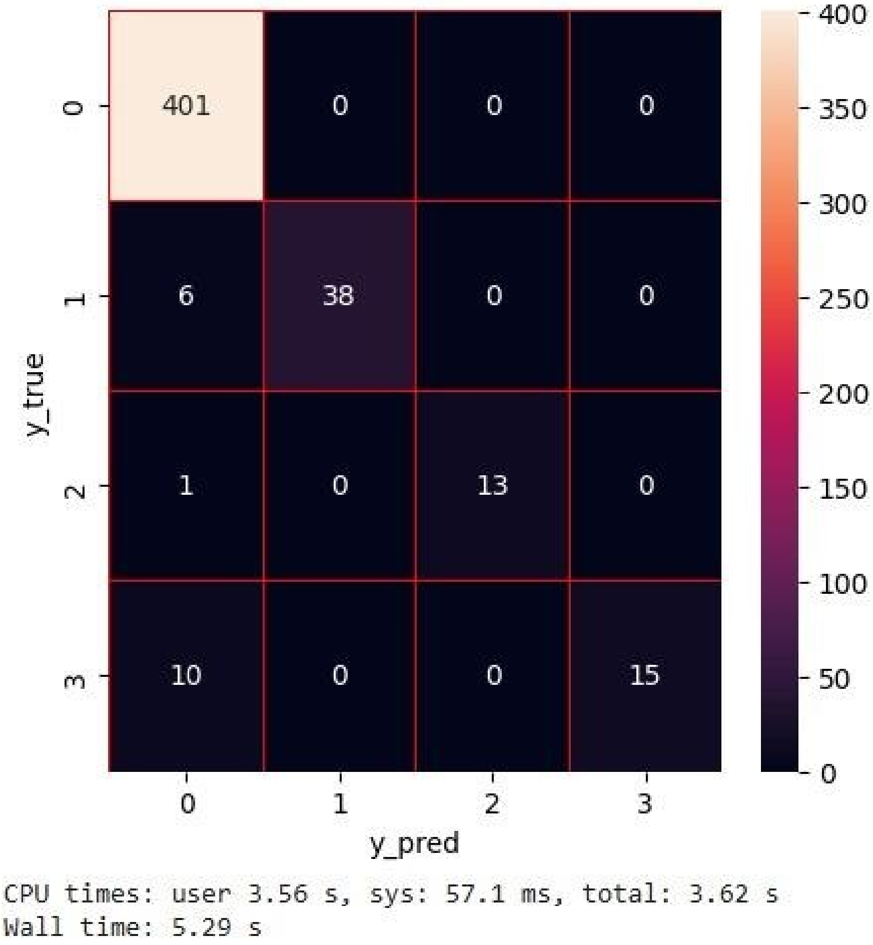
4.2 Model results

4.2.1. Classifier-1: Random Forest



```
[[401   0   0   0]
 [  6  38   0   0]
 [  1   0  13   0]
 [ 10   0   0  15]]

Accuracy: 0.96

Micro Precision: 0.96
Micro Recall: 0.96
Micro F1-score: 0.96

Macro Precision: 0.99
Macro Recall: 0.85
Macro F1-score: 0.90

Weighted Precision: 0.97
Weighted Recall: 0.96
Weighted F1-score: 0.96

Classification Report

              precision    recall  f1-score   support

     Class 1       0.96      1.00      0.98       401
     Class 2       1.00      0.86      0.93        44
     class 3       1.00      0.93      0.96        14
     class 4       1.00      0.60      0.75        25

    accuracy                           0.96       484
   macro avg       0.99      0.85      0.90       484
weighted avg       0.97      0.96      0.96       484
```

4.2.1.a Classifier-1 Evaluation Metrics

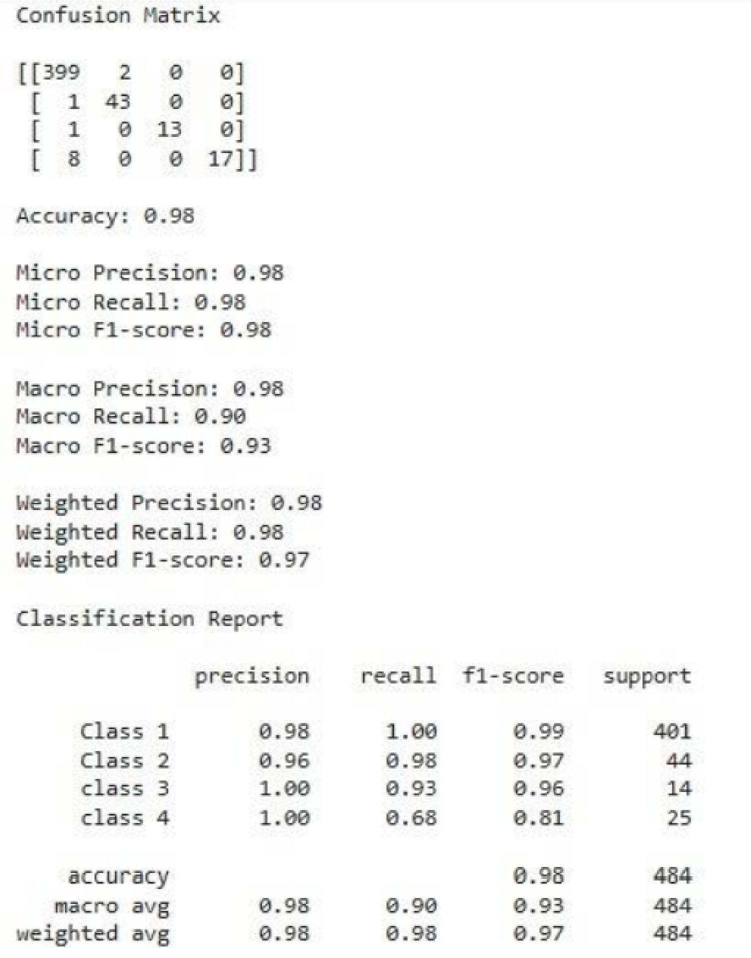


```
CPU times: user 3.56 s, sys: 57.1 ms, total: 3.62 s
Wall time: 5.29 s
```

4.2.1.b Classifier-1 Heat Map

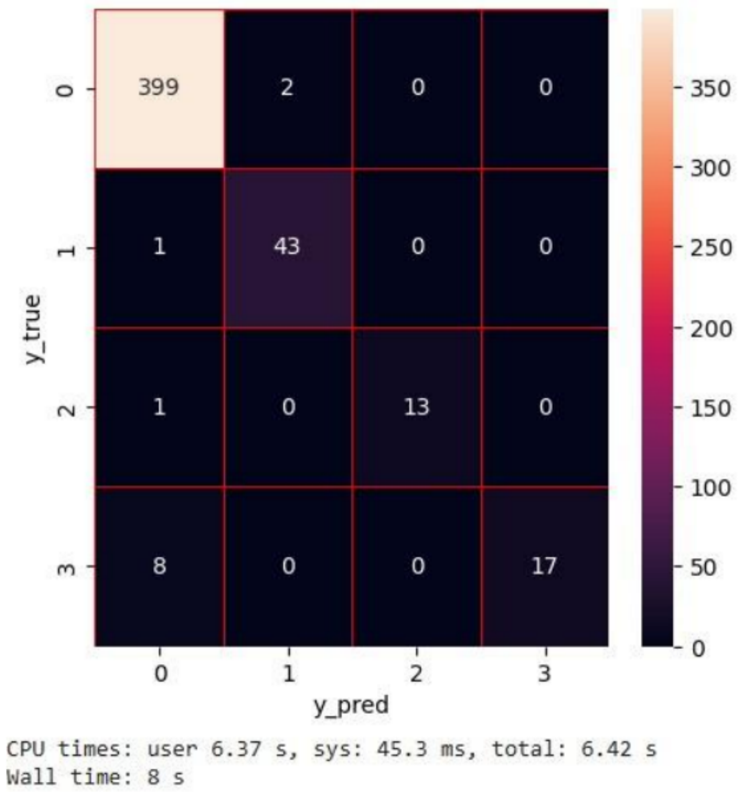
i. Random Forest gives Accuracy (0.96), Precision (0.97), Recall (0.96), F1-Score (0.96)

4.2.2. Classifier-2: XG Boost



```
Confusion Matrix

[[399   2   0   0]
 [  1  43   0   0]
 [  1   0  13   0]
 [  8   0   0  17]]

Accuracy: 0.98

Micro Precision: 0.98
Micro Recall: 0.98
Micro F1-score: 0.98

Macro Precision: 0.98
Macro Recall: 0.90
Macro F1-score: 0.93

Weighted Precision: 0.98
Weighted Recall: 0.98
Weighted F1-score: 0.97

Classification Report

              precision    recall  f1-score   support

     Class 1       0.98      1.00      0.99       401
     Class 2       0.96      0.98      0.97        44
     class 3       1.00      0.93      0.96        14
     class 4       1.00      0.68      0.81        25

    accuracy                           0.98       484
   macro avg       0.98      0.90      0.93       484
weighted avg       0.98      0.98      0.97       484
```

4.2.2.a Classifier-2 Evaluation Metrics values



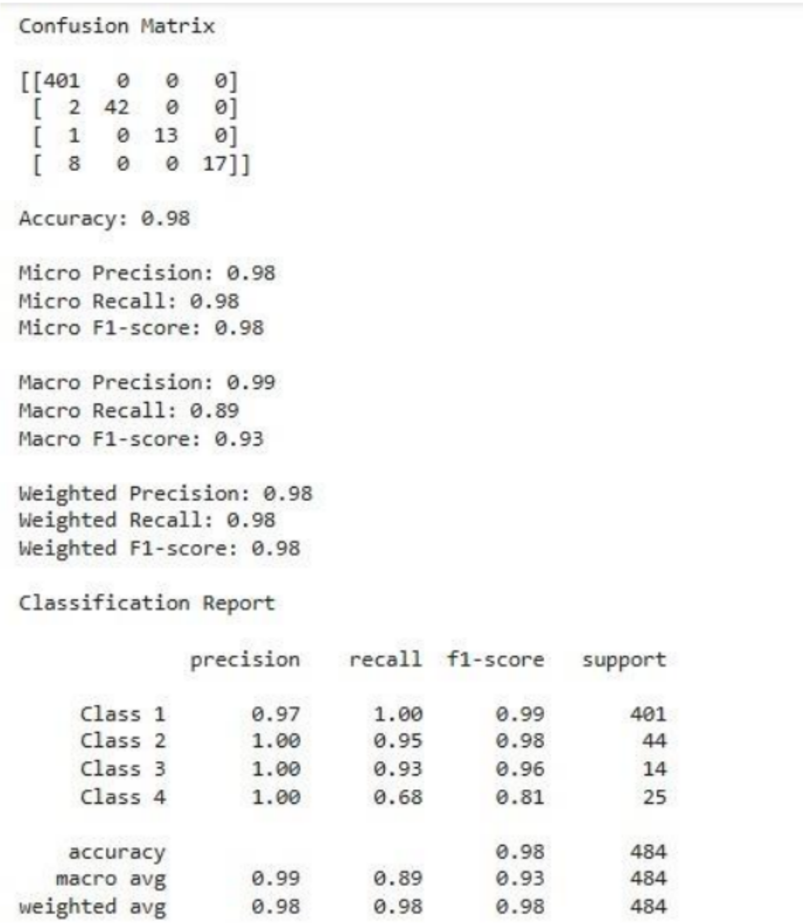
	0	1	2	3
0	399	2	0	0
1	1	43	0	0
2	1	0	13	0
3	8	0	0	17

```
CPU times: user 6.37 s, sys: 45.3 ms, total: 6.42 s
Wall time: 8 s
```

4.2.2.b. Classifier-2 Heat Map

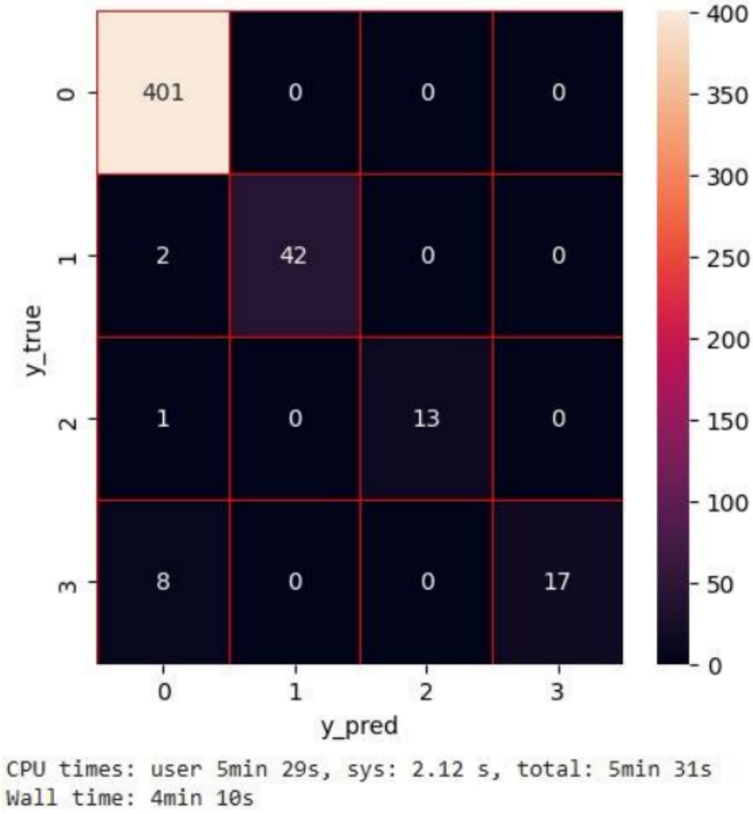
ii. XGBoost gives the Accuracy (0.98), Precision (0.98), Recall (0.98), F1-Score (0.98)

4.2.3. Classifier-3: CatBoost



```
Confusion Matrix

[[401   0   0   0]
 [  2  42   0   0]
 [  1   0  13   0]
 [  8   0   0  17]]

Accuracy: 0.98

Micro Precision: 0.98
Micro Recall: 0.98
Micro F1-score: 0.98

Macro Precision: 0.99
Macro Recall: 0.89
Macro F1-score: 0.93

Weighted Precision: 0.98
Weighted Recall: 0.98
Weighted F1-score: 0.98

Classification Report

              precision    recall  f1-score   support

     Class 1       0.97      1.00      0.99       401
     Class 2       1.00      0.95      0.98        44
     Class 3       1.00      0.93      0.96        14
     Class 4       1.00      0.68      0.81        25

    accuracy                           0.98       484
   macro avg       0.99      0.89      0.93       484
weighted avg       0.98      0.98      0.98       484
```

4.2.3.a. Classifier-3 Evaluation Metrics values

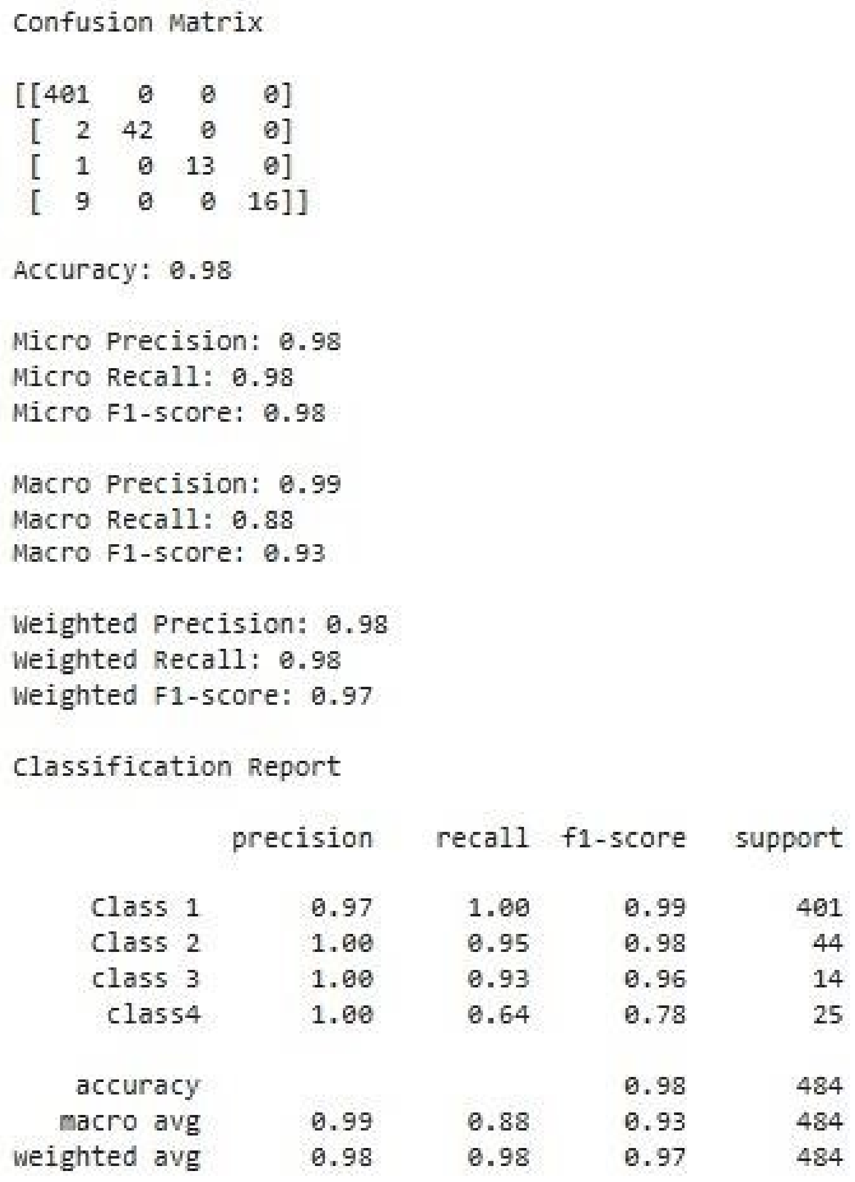


```
CPU times: user 5min 29s, sys: 2.12 s, total: 5min 31s
Wall time: 4min 10s
```

4.2.3.b. Classifier-3 Heat Map

i. CatBoost gives the Accuracy (0.98), Precision(0.98), Recall(0.98),F1-Score (0.98)

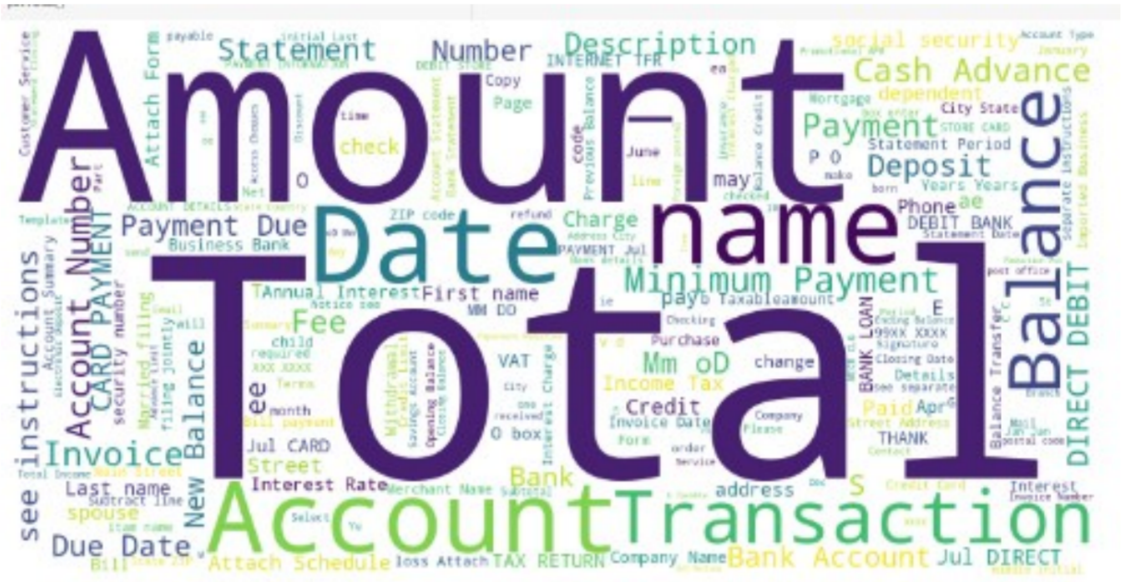
4.2.4. Classifier-4: Voting Classifier



```
Confusion Matrix

[[401   0   0   0]
 [  2  42   0   0]
 [  1   0  13   0]
 [  9   0   0  16]]

Accuracy: 0.98

Micro Precision: 0.98
Micro Recall: 0.98
Micro F1-score: 0.98

Macro Precision: 0.99
Macro Recall: 0.88
Macro F1-score: 0.93

Weighted Precision: 0.98
Weighted Recall: 0.98
Weighted F1-score: 0.97

Classification Report

              precision    recall  f1-score   support

     Class 1       0.97      1.00      0.99       401
     Class 2       1.00      0.95      0.98        44
     class 3       1.00      0.93      0.96        14
      class4       1.00      0.64      0.78        25

    accuracy                           0.98       484
   macro avg       0.99      0.88      0.93       484
weighted avg       0.98      0.98      0.97       484
```

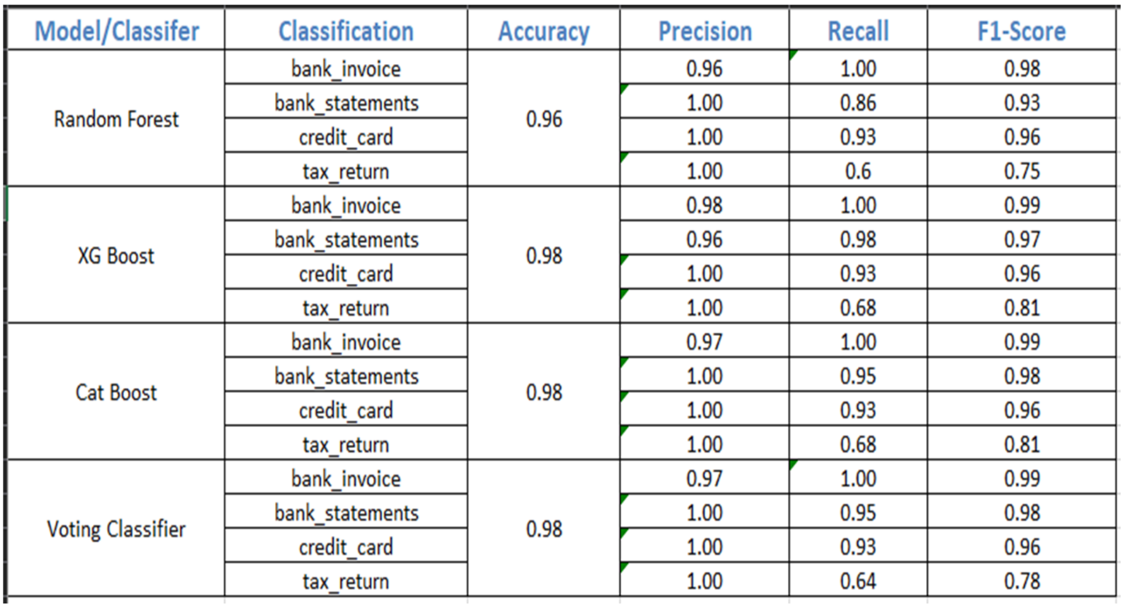
4.2.4.a. Classifier-4 Evaluation Metrics values

4.3. Word Cloud



4.3.a. Word Cloud Representation of Keywords

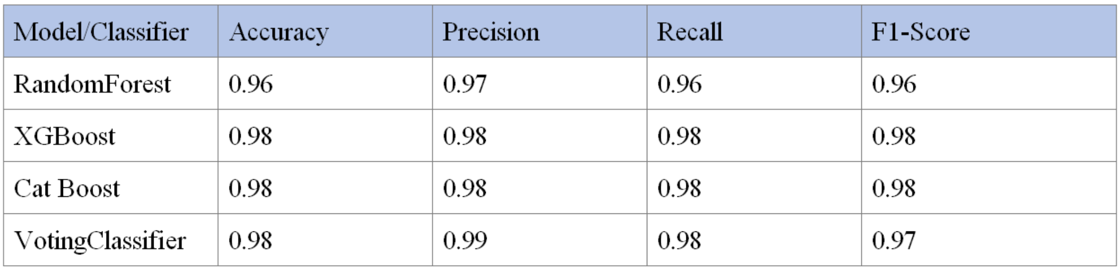
4.4. Comparative Analysis of Different Classifier's Results



Model/Classifer	Classification	Accuracy	Precision	Recall	F1-Score
Random Forest	bank_invoice	0.96	0.96	1.00	0.98
	bank_statements		1.00	0.86	0.93
	credit_card		1.00	0.93	0.96
	tax_return		1.00	0.6	0.75
XG Boost	bank_invoice	0.98	0.98	1.00	0.99
	bank_statements		0.96	0.98	0.97
	credit_card		1.00	0.93	0.96
	tax_return		1.00	0.68	0.81
Cat Boost	bank_invoice	0.98	0.97	1.00	0.99
	bank_statements		1.00	0.95	0.98
	credit_card		1.00	0.93	0.96
	tax_return		1.00	0.68	0.81
Voting Classifier	bank_invoice	0.98	0.97	1.00	0.99
	bank_statements		1.00	0.95	0.98
	credit_card		1.00	0.93	0.96
	tax_return		1.00	0.64	0.78

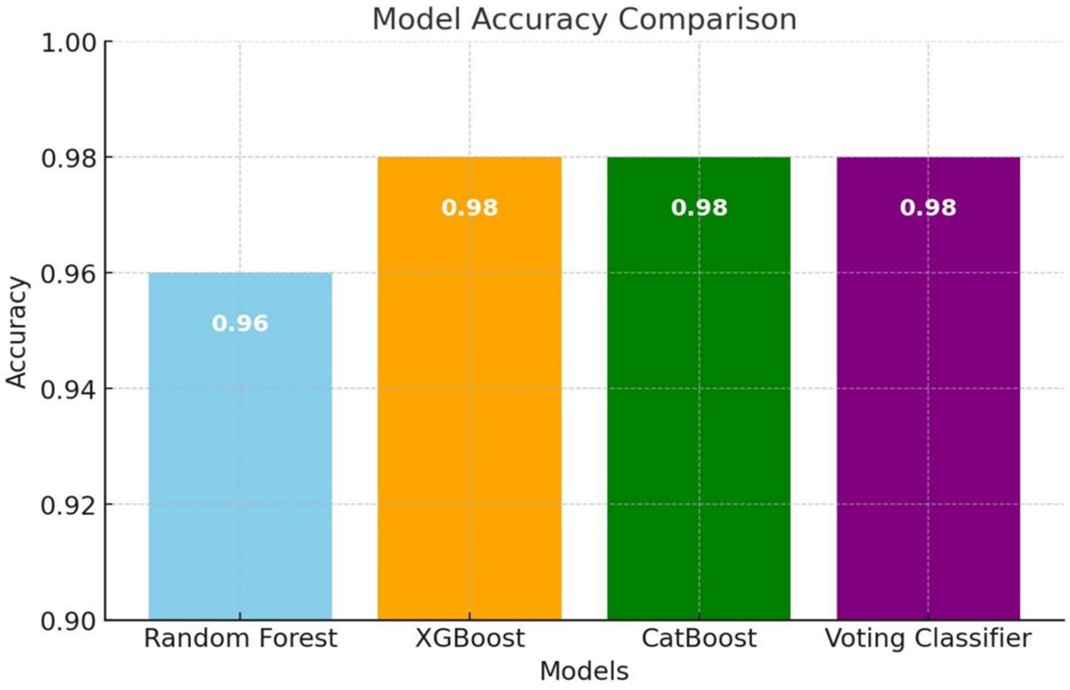
4.4.a. Comparative Analysis of Evaluation Metrics

4.5. Accuracy Analysis of Algorithms



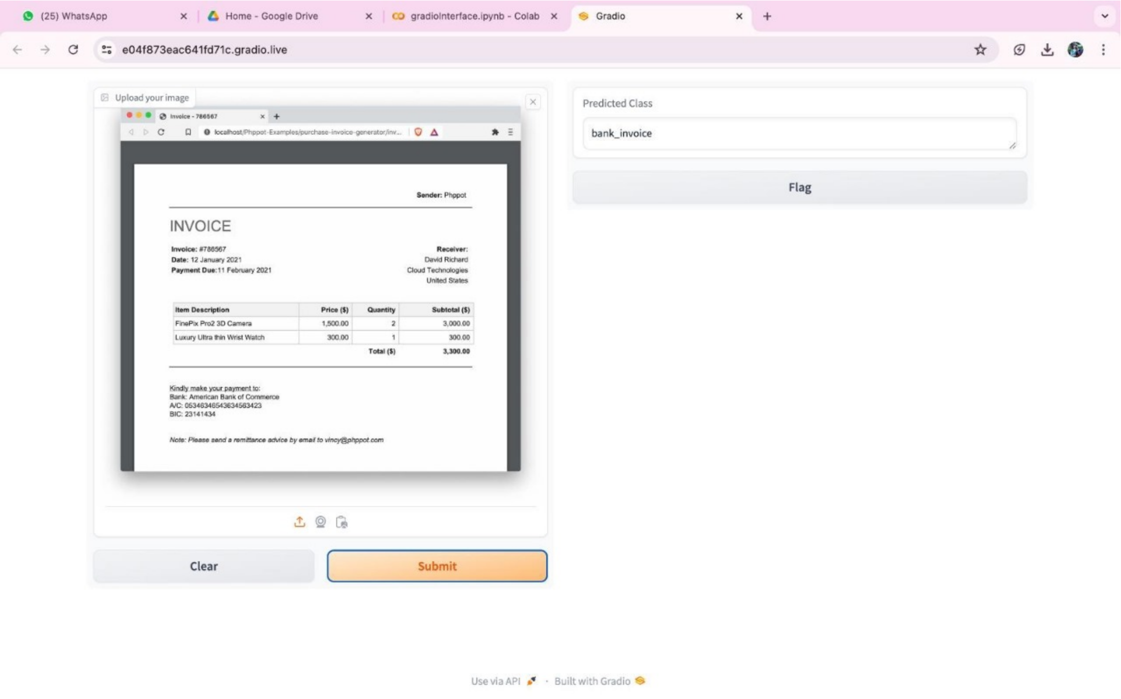
Model/Classifier	Accuracy	Precision	Recall	F1-Score
RandomForest	0.96	0.97	0.96	0.96
XGBoost	0.98	0.98	0.98	0.98
Cat Boost	0.98	0.98	0.98	0.98
VotingClassifier	0.98	0.99	0.98	0.97

4.5.a. Accuracy Analysis



4.5.b. Accuracy Analysis of Classifiers

4.6. Results


(25) WhatsApp
Home - Google Drive
gradioInterface.ipynb - Colab
Gradio
e04f873eac641fd71c.gradio.live
Upload your image
Invoice - 786567
localhost/Phppot-Examples/purchase-invoice-generator/inv...
Clear
Submit
Flag
Use via API · Built with Gradio

Sender: Phppot

INVOICE

Invoice: #786567
Date: 12 January 2021
Payment Due:11 February 2021

Receiver:
David Richard
Cloud Technologies
United States

Item Description	Price ($)	Quantity	Subtotal ($)
FinePix Pro2 3D Camera	1,500.00	2	3,000.00
Luxury Ultra thin Wrist Watch	300.00	1	300.00
		Total ($)	**3,300.00**

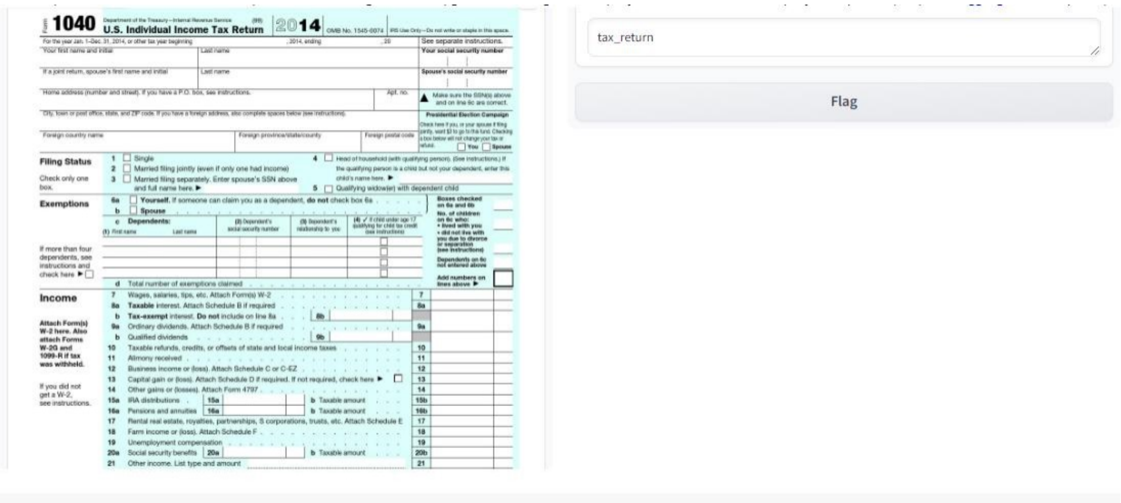
Kindly make your payment to:
Bank: American Bank of Commerce
A/C: 05346346543634563423
BIC: 23141434

Note: Please send a remittance advice by email to vincy@phppot.com

Predicted Class

bank_invoice

4.6.a. Invoice



Form **1040** Department of the Treasury—Internal Revenue Service (99) **U.S. Individual Income Tax Return** 2014 OMB No. 1545-0074 IRS Use Only—Do not write or staple in this space.

For the year Jan. 1–Dec. 31, 2014, or other tax year beginning , 2014, ending , 20 See separate instructions.

Your first name and initial | Last name | **Your social security number**

If a joint return, spouse's first name and initial | Last name | **Spouse's social security number**

Home address (number and street). If you have a P.O. box, see instructions. | Apt. no. | ▲ Make sure the SSN(s) above and on line 6c are correct.

City, town or post office, state, and ZIP code. If you have a foreign address, also complete spaces below (see instructions).

Presidential Election Campaign
Check here if you, or your spouse if filing jointly, want $3 to go to this fund. Checking a box below will not change your tax or refund. □ You □ Spouse

Foreign country name | Foreign province/state/county | Foreign postal code

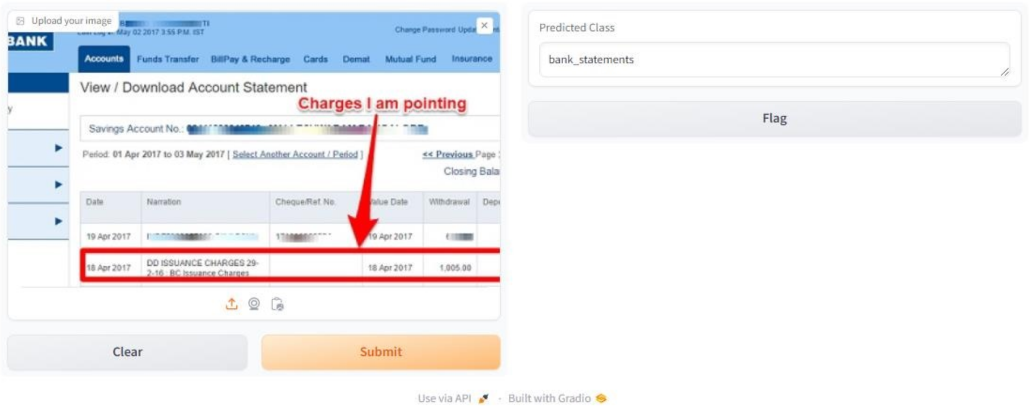
Filing Status
Check only one box.
1 □ Single
2 □ Married filing jointly (even if only one had income)
3 □ Married filing separately. Enter spouse's SSN above and full name here. ▶
4 □ Head of household (with qualifying person). (See instructions.) If the qualifying person is a child but not your dependent, enter this child's name here. ▶
5 □ Qualifying widow(er) with dependent child

Exemptions
6a □ **Yourself.** If someone can claim you as a dependent, **do not** check box 6a
b □ **Spouse**
c **Dependents:** (1) First name Last name | (2) Dependent's social security number | (3) Dependent's relationship to you | (4) ✓ if child under age 17 qualifying for child tax credit (see instructions)
If more than four dependents, see instructions and check here ▶ □
d Total number of exemptions claimed
Boxes checked on 6a and 6b
No. of children on 6c who:
• lived with you
• did not live with you due to divorce or separation (see instructions)
Dependents on 6c not entered above
Add numbers on lines above ▶

Income
Attach Form(s) W-2 here. Also attach Forms W-2G and 1099-R if tax was withheld.
If you did not get a W-2, see instructions.
7 Wages, salaries, tips, etc. Attach Form(s) W-2 7
8a **Taxable** interest. Attach Schedule B if required 8a
b **Tax-exempt** interest. **Do not** include on line 8a 8b
9a Ordinary dividends. Attach Schedule B if required 9a
b Qualified dividends 9b
10 Taxable refunds, credits, or offsets of state and local income taxes 10
11 Alimony received 11
12 Business income or (loss). Attach Schedule C or C-EZ 12
13 Capital gain or (loss). Attach Schedule D if required. If not required, check here ▶ □ 13
14 Other gains or (losses). Attach Form 4797 14
15a IRA distributions 15a b Taxable amount 15b
16a Pensions and annuities 16a b Taxable amount 16b
17 Rental real estate, royalties, partnerships, S corporations, trusts, etc. Attach Schedule E 17
18 Farm income or (loss). Attach Schedule F 18
19 Unemployment compensation 19
20a Social security benefits 20a b Taxable amount 20b
21 Other income. List type and amount 21

tax_return

4.6.b. Tax-Return



[redacted]
May 02 2017 3:55 P.M. IST

View / Download Account Statement

Savings Account No.: [redacted]

Period: 01 Apr 2017 to 03 May 2017 [Select Another Account / Period]

<< Previous Page

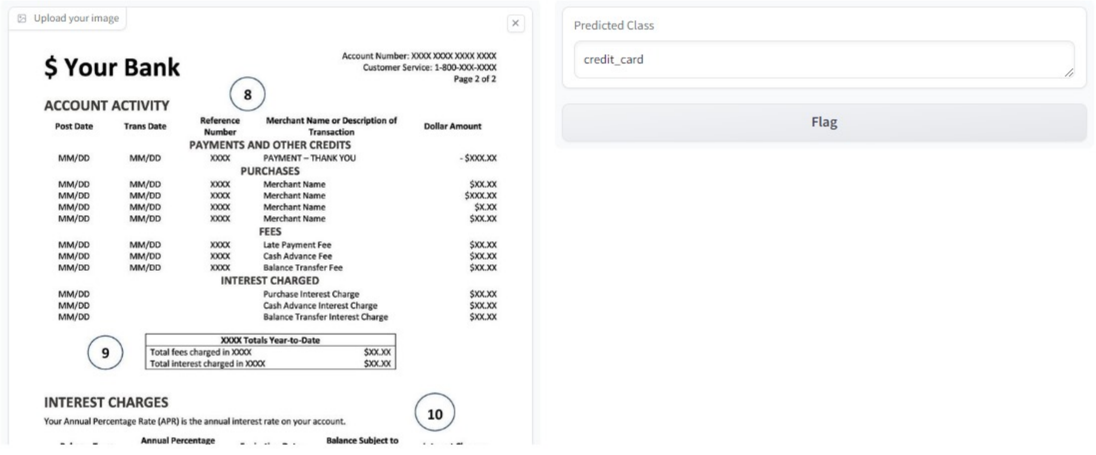
Closing Bala

Date	Narration	Cheque/Ref No.	Value Date	Withdrawal	Depo
19 Apr 2017	[redacted]	[redacted]	19 Apr 2017	[redacted]	
18 Apr 2017	DD ISSUANCE CHARGES 29-2-16 : BC Issuance Charges		18 Apr 2017	1,005.00	

Predicted Class

bank_statements

4.6.c. Bank-Statement



$ Your Bank

Account Number: XXXX XXXX XXXX XXXX
Customer Service: 1-800-XXX-XXXX
Page 2 of 2

ACCOUNT ACTIVITY

Post Date	Trans Date	Reference Number	Merchant Name or Description of Transaction	Dollar Amount
			PAYMENTS AND OTHER CREDITS	
MM/DD	MM/DD	XXXX	PAYMENT – THANK YOU	- $XXX.XX
			PURCHASES	
MM/DD	MM/DD	XXXX	Merchant Name	$XX.XX
MM/DD	MM/DD	XXXX	Merchant Name	$XXX.XX
MM/DD	MM/DD	XXXX	Merchant Name	$X.XX
MM/DD	MM/DD	XXXX	Merchant Name	$XX.XX
			FEES	
MM/DD	MM/DD	XXXX	Late Payment Fee	$XX.XX
MM/DD	MM/DD	XXXX	Cash Advance Fee	$XX.XX
MM/DD	MM/DD	XXXX	Balance Transfer Fee	$XX.XX
			INTEREST CHARGED	
MM/DD			Purchase Interest Charge	$XX.XX
MM/DD			Cash Advance Interest Charge	$XX.XX
MM/DD			Balance Transfer Interest Charge	$XX.XX

XXXX Totals Year-to-Date	
Total fees charged in XXXX	$XX.XX
Total interest charged in XXXX	$XX.XX

INTEREST CHARGES

Your Annual Percentage Rate (APR) is the annual interest rate on your account.

Annual Percentage | Balance Subject to

Predicted Class

credit_card

4.6.d. Credit-Card

5. OBSERVATIONS

1. It is observed that the classification of bank documents achieved better results with different ensemble Machine Learning models (i.e., Random Forest, XGBoost, and CatBoost), and the Voting Classifier is considered the benchmark model.

2. The performance of individual classifiers is measured based on the output values of evaluation metrics (Accuracy, Precision, Recall, and F1-Score):

   2.1 **Random Forest** – Accuracy (0.96), Precision (0.97), Recall (0.96), F1-Score (0.96).

   2.2 **XGBoost** – Accuracy (0.98), Precision (0.98), Recall (0.98), F1-Score (0.98).

   2.3 **CatBoost** – Accuracy (0.98), Precision (0.98), Recall (0.98), F1-Score (0.98).

   2.4 **Voting Classifier** – Accuracy (0.98), Precision (0.99), Recall (0.98), F1-Score (0.97).

Based on the results obtained from the above-mentioned classifiers, it is observed that the Voting Classifier, XGBoost, and CatBoost algorithms achieved the highest accuracy (98%) and performed equally well among all four classifiers.

6. CONCLUSION

Document Classification is an essential task in the banking system for organizing and managing large amounts of data. The proposed system uses advanced ensemble machine learning algorithms and natural language processing techniques, which help to automatically categorize and classify bank documents that contain both structured and unstructured data. The salient features extracted from the ensemble machine learning model are fed into output classification. The proposed ensemble machine learning model (Voting Classifier) achieved an accuracy of 98% in comparison with Random Forest (96%), XGBoost (98%), and CatBoost (98%).

The process of classification of bank documents helps banks to improve their efficiency, reduce operating costs, and improve customer service.

7. FUTURE WORK

As part of future work, the banking system needs to extract relevant information from the documents. Automatic data extraction that automatically extracts the data from different types of documents can save time and reduce errors. Finance and banking sectors are tremendously developing globally. Integrating Deep Learning and Natural Language Processing techniques into document classification can enhance accuracy and efficiency by reducing errors.

8. REFERENCES

1. Sarosh Dandoti, (2022), *Text Document Classification System*, International Research Journal of Engineering and Technology (IRJET), 9(6), 2847-2850.

2. Arslan, Ö., & Uymaz, S. A. (2022).*Classification of Invoice Images by Convolutional Neural Networks.* Journal of Advanced Research in Natural and Applied Sciences, 8(1), 8-25.

3. Sheth, V., Tripathi, U., & Sharma, A. (2022). *A Comparative Analysis of Machine Learning Algorithms for Classification Purpose.* Procedia Computer Science, 215, 422-431.

4. Ghumade, T. G., & Deshmukh, R. A. (2019). *A document classification using NLP and recurrent neural network.* Int. J. Eng. Adv. Technol, 8(6), 632-636.

5. Engin, D., Emekligil, E., Oral, B., Arslan, S., & Akpınar, M. (2019). *Multimodal deep neural networks for banking document classification.* In International Conference on Advances in Information Mining and Management (pp. 21-25).

6. Trivedi, S., Malawat, K., Yerawar, N., Chaudhary, S., & Kamble, A. (2020). *Document Classification Using Machine Learning.* 21(4), 6-13.

7. Engin, D., Emekligil, E., Oral, B., Arslan, S., & Akpınar, M. (2019). *Multimodal deep neural networks for banking document classification.* In International Conference on Advances in Information Mining and Management (pp. 21-25).

8. Sharma, K., Gaikwad, A., Patil, S., Kumar, P., & Salapurkar, D. P. (2018). *Automated Document Summarization and Classification Using DL.* International Research Journal of Engineering and Technology, 5(06).

9. Guibin Chen, Deheng, Zhenchang Xing (2017). *Ensemble Application of Convolutional and Recurrent Neural Networks for Multi-label Text Categorization.* IEEE.

10. Md. Saiful Islam, (2017). *A Support Vector Machine Mixed with TF-IDF Algorithm to Categorize Document.* IEEE.

11. Cüceloğlu, İ., & Oğul, H. *Automated categorization of scanned bank documents from extracted text content.*

12. Basarkar, A. (2017). *Document classification using machine learning.*

13. Yaram, S. (2016, August). *Machine learning algorithms for document clustering and fraud detection.* In 2016 International Conference on Data Science and Engineering (ICDSE) (pp. 1–6). IEEE.

14. Ting, S. L., Ip, W. H., & Tsang, A. H. (2011). *Is Naive Bayes a good classifier for document classification.* International Journal of Software Engineering and Its Applications, 5(3), 37–46.

15. A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks.* In Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

16. K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014.

17. Szegedy et al., *Going deeper with convolutions.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

18. K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

19. M. Z. Afzal, A. Kölsch, S. Ahmed and M. Liwicki,*Cutting the error by half:Investigation of very deep CNN and advanced training strategies for document image classification.*