Journal of Basics and Applied Sciences Research (JOBASR)
ISSN (print): 3026-9091, ISSN (online): 1597-9962
Volume 4(3) May 2026
DOI: https://dx.doi.org/10.4314/jobasr.v4i3.20

XGBoost-Based Supervised Machine Learning for Automated Tag Prediction in Digital File Organization: An Explainable AI Approach

Aisha Musa Kaita[1*], Saleh El-Yakub Abdullahi[2] & Bilkisu Muhammad-Bello[3]
[1,3]Department of Software Engineering, Nile university of Nigeria, Abuja. Nigeria
[2]Department of Computer Science, Nile university of Nigeria, Abuja, Nigeria
[*]Corresponding Author Email: ayshamakaita@gmail.com

ABSTRACT
Due to the exponential growth in digital data, the process of manual tagging of files is no longer sufficient to achieve efficiency in the retrieval process. Therefore, this study proposes a supervised machine learning model using the Extreme Gradient Boosting (XGBoost) algorithm with the Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction method, which incorporates unigrams and bigrams, to predict document tags. The model is tested using two different datasets, namely the Biomedical Text Publication Classification data set with 7,569 instances and the BBC News Articles dataset with 2,225 instances. Additionally, a fivefold stratified cross-validation method was used to evaluate the performance of the proposed model and mitigate the potential biases. Explanations have been provided using methods such as LIME and SHAP. Using the Extreme Gradient Boosting classifier, an accuracy score of 99.74% was achieved using the biomedical dataset, and 94% accuracy was achieved when training the model using the news article dataset. In addition, the model performed better than baselines, including Support Vector Machine (SVM) and Random ForestThe existing automated tag generation algorithms lack the ability to generate results quickly due to high computational costs, and are often not easily interpretable. This study proposes a framework that uses a combination of the XGBoost and TF-IDF models, along with explainable AI techniques such as SHAP and LIME, to provide transparent predictions of tags.

Keywords:
XGBoost,
SHAP,
LIME,
Digital file organization,
Natural language
Processing

INTRODUCTION

There has been an explosion in size in the digital information ecosystem. The world's total volume of information has reached up to 149 zettabytes by 2024. For the year 2025, the predicted volume has reached up to 181 zettabytes or above (Mgoldring, 2024). In this era of digital information environment, there has been emergence of digital file management as one of the most crucial problems. The use of a traditional approach that involves hierarchical folders for organizing documents cannot be used for managing digital information because such a process compels users to place their information into a single category whereas in actuality, it can belong to different categories (Henderson & Srinivasan, 2022). Users of digital information are therefore experiencing what can be referred to as the 'filing dilemma' as coined by Kim & Lee (2023).

Here, tagging mechanisms have been put forward as an alternative approach that is more flexible and multi-dimensional than hierarchical systems, where tags can be attached to files on the basis of their content in multiple ways (Wang et al., 2023). Despite being more effective than hierarchical categorization, manual tagging has faced several issues such as inconsistency, lack of vocabulary coherence, cold start issue related to newly generated files, and time-consuming procedures (Bansal et al., 2025). As a result, current studies have emphasized the use of automated tagging systems using AI.

The existing approaches to solving the problem of automatic tagging have several limitations. For instance, the LLM-based approaches despite their high accuracy, incur significant costs and, therefore, are impractical for use in enterprise environments (Liu et al., 2024). Another group of approaches, multimodal approaches such as FileScope, although quite accurate (92.3%), show degraded performance when encountering corrupted input (Wang et al., 2024). Almost all the approaches mentioned above, as can be expected due to their reliance on AI,

How to cite this article: Aisha M. K., Saleh E. A. & Bilkisu M. B. (2026). XGBoost-Based Supervised Machine Learning for Automated Tag Prediction in Digital File Organization: An Explainable AI Approach. *Journal of Basics and Applied Sciences Research*, *4*(3), 178-189. https://dx.doi.org/10.4314/jobasr.v4i3.20

operate as a "black box" without providing any justification behind tags generated by those approaches (Johnson et al., 2024).

In order to overcome such limitations, the current paper suggests an accurate, but efficient model of tag prediction based on XGBoost algorithm along with Explainable Artificial Intelligence (XAI). According to the suggested scheme, the TF-IDF vectorization with unigrams and bigrams will be utilized to encode the features, while the algorithms of LIME and SHAP will be used to achieve local/global explainability. The suggested model will be tested using two benchmark data sets, one dedicated to biomedical science papers and the other to news articles. Despite receiving considerable attention, there remain some deficiencies in the existing literature on AI-based methods for organizing files as follows: high computational expense (Liu et al., 2024), low interpretability (Johnson et al., 2024), and inadequate validation methods for assessing reliability. The research at hand contributes to addressing these deficiencies through combining XGBoost with TF-IDF features, using 5-fold stratified cross-validation and employing SHAP/LIME for interpretability.

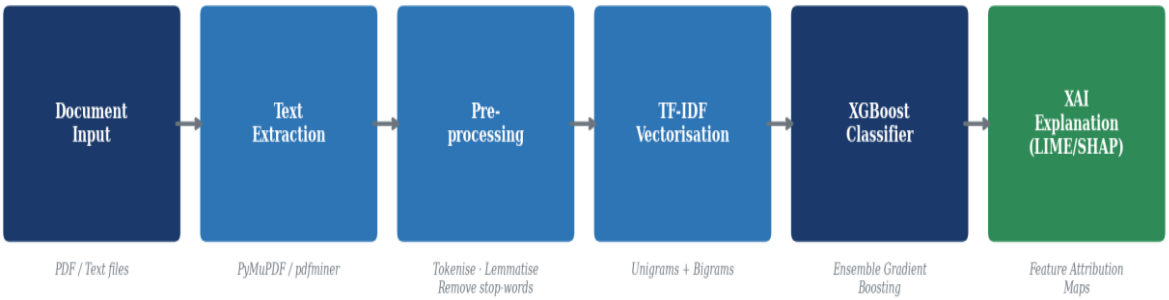

Figure 1. *End-to-end tag prediction pipeline — from raw document input through text extraction, preprocessing, TF-IDF vectorization, XGBoost classification, to XAI explanation output*.

The rest of this paper is organized as follows: Section 2 gives a general overview of the literature related to this research, Section 3 explains the proposed method in detail, Section 4 and 5 show and discuss the experimental results, and finally Section 5 concludes this research with its implications and future research directions.

Firstly, the objective of this research is to (i) build a light-weight XGBoost-TF-IDF model capable of automatically predicting tags in documents; (ii) test its efficacy using 5-fold stratified cross-validation on two distinct benchmark datasets; and (iii) assess its interpretability using explainable techniques (SHAP and LIME). The major premise of this work is that the use of an ensemble-based non-linear classifier such as XGBoost and n-grams TF-IDF features will yield better results compared to linear classifiers (e.g., SVM and RF) in multi-class text classification problems. The choice of using XGBoost is driven by factors such as: (1) its ability to strike a balance between accuracy and computational speed; (2) having regularization capability which prevents overfitting; and (3) being built into SHAP explainability methods. The main contributions of this research will be: (1) building a validated and explainable machine learning model; (2) cross-domain evaluation of XGBoost on biomedical and news datasets; and (3) showing empirically that the proposed framework can achieve SOTA performance while remaining computationally efficient compared to transformers and LLMs.

Information Organization and the Limits of Traditional Systems

Information classification systems have traditionally been shaped by the cultural and academic milieu in which they were created (Hjørland, 2014). The Dewey Decimal Classification System and Library of Congress were preeminent in library science for over a century but were created in an era of stable, hierarchical knowledge structures (Makwana, 2024). However, as information in the digital world is fluid, diverse, and ever-changing, the inflexibility of hierarchical classification is proving inadequate. Faceted classification theory was first proposed by Morrison & Schyns in 2001, allowing multi-dimensional classification by simultaneously classifying an object by various attributes, an ideal application of digital tagging systems (Gottschewski-Meyer et al., 2024).

Folksonomy, Social Tagging, and Collaborative Systems

The rise of Web 2.0 technologies has also led to the adoption of collaborative tagging, also known as folksonomy. In this regard, Halpin (2012) noted that collaborative tagging systems tend to become more stable as natural consensus-based vocabularies emerge. However, Efe (2021) and Gupta et al. (2011) also noted that synonymy proliferation, homonymy ambiguity, and tag granularity inconsistency have been some of the challenges associated with collaborative tagging. These challenges have been formally defined by Bansal et al. (2025) as three structural challenges associated with collaborative tagging: the cold start problem, inconsistency in tag quality, and confusion due to polysemous terms.

Content-Based and Machine Learning Approaches to Tag Generation

For instance, the statistical approach for the generation of tags has relied on the application of TF-IDF, which has been used to identify terms that are both common within a particular document and, at the same time, scarce within the entire corpus (Nafis & Awang, 2021). However, Fioravanti & Siyanova-Chanturia (2024) were able to show how the application of keyphrase extraction resulted in the extraction of more information compared to the extraction of single-word terms. The application of the combination of linguistic and statistical approaches, such as part-of-speech tagging and noun phrase extraction, was also found to result in better keyword extraction, as was demonstrated by Firoozeh et al. (2020). The application of graph-based approaches, such as TextRank, has also resulted in the extraction of keywords based on the context (Mihalcea & Hassan, 2024).

Machine learning techniques have reformulated the problem of tag prediction as a multi-label classification problem. Cakar et al. (2024) have also evaluated binary relevance, label powerset, and k-nearest neighbor for predicting movie tags based on BERT-based features. Deep learning models such as Convolution Neural Networks (CNNs) and Residual Networks (ResNets) have greatly contributed to the improvement of image tagging by taking into account hierarchical representations of the data (Khan et al., 2020; Duta et al., 2021). In the case of text data, transformer-based models such as BERT have shown state-of-the-art results for a variety of natural language processing tasks (Shreyashree et al., 2022). However, Jiang et al. (2023) have also shown that sentence-based models are not effective for informal and social media-based text data.

Explainable AI in Document Classification

One of the major challenges facing the implementation of AI-based tagging systems is the black box problem. The black box problem is a problem of a lack of understanding and auditability of the underlying process. Explainable AI techniques have been developed to overcome this problem. Techniques such as LIME have been developed for the creation of explanations that are locally faithful to individual predictions by the use of surrogate models (Ribeiro et al., 2016). Techniques such as SHAP have been developed for the creation of explanations that are theoretically sound by the use of Shapley values for ensuring the accuracy of feature importance estimation. The use of explainable AI in document classification has been recognized as a key requirement for the implementation of trustworthy AI in information management systems (Johnson et al., 2024).

XGBoost for Text Classification

XGBoost is an efficient and scalable gradient-boosting framework that is suitable for distributed computing. This algorithm uses an ensemble of decision trees in a sequential fashion, where each tree attempts to correct the errors made by the previous tree. This algorithm always yields competitive results in classification problems involving structured data. Additionally, it is computationally efficient due to parallel processing capabilities. Zdravkovic (2021) proved that supervised machine learning with TF-IDF features is an effective solution in multi-class text classification problems in scientific texts, where SVM yields the best results. In this study, it is hypothesized that an ensemble of decision trees using the non-linear XGBoost algorithm will yield better results compared to the linear SVM algorithm in text classification problems in technical texts as well as general texts.

Research Gap

Although substantial progress has been made in the development of efficient automatic document tagging and classification, some fundamental research gaps have to be filled in the literature. Firstly, the majority of the high-accuracy approaches, including the use of LLM-based taggers, are computationally infeasible for practical applications in resource-constrained settings (Liu et al., 2024). Secondly, the application of the ensemble machine learning approach together with the use of XAI for tag prediction in the overall digital file management workflow has received little research attention. Finally, the cross-domain evaluation of a single framework for biomedical and general news domain corpora has received little research attention. The current research directly addresses the above research gaps by proposing a novel efficient, transparent, and domain-generalizable XGBoost-TF-IDF-based framework with the use of LIME and SHAP for tag prediction.

MATERIALS AND METHODS

Research Design and System Architecture

A quantitative experimental research design was adopted, and two heterogeneous benchmark datasets were evaluated under identical experimental conditions. The architecture of the system consists of: document input, text extraction/preprocessing, feature engineering using TF-IDF, classification using three different models, application of 5-fold stratified cross-validation, and finally, model explanation using LIME and SHAP techniques. Figure 2 shows the detailed architecture.

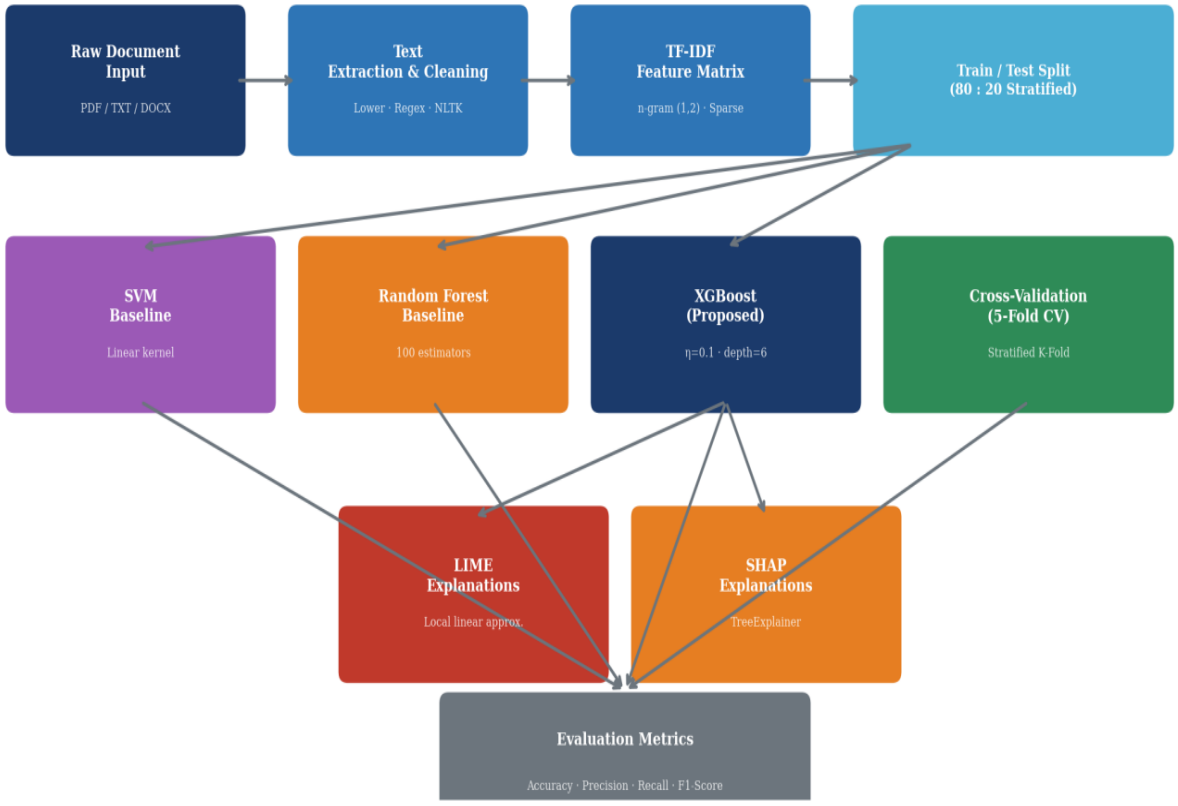

Figure 2. Detailed system architecture showing the cross-validation workflow alongside the training and test evaluation pipeline. SVM, Random Forest, and XGBoost are evaluated under identical conditions.

Datasets

Two benchmark datasets were used to measure the performance of the model in different and contrasting textual domains:

The two datasets were deliberately selected to provide contrasting textual domains — specialized biomedical vocabulary versus general-domain journalism — enabling evaluation of the framework's cross-domain generalizability. Both datasets are publicly available on the Kaggle platform, are widely used in the text classification literature, and do not contain personally identifiable information; thus, no ethical approval was required for their use. All experiments were implemented in Python 3.10 using the scikit-learn library (v1.3.0), XGBoost (v2.0.0), SHAP (v0.43.0), and LIME (v0.2.0.1) on a standard workstation (Intel Core i7-12700H, 16 GB RAM). A random seed of 42 was set throughout all experiments to ensure reproducibility. Hyperparameters for XGBoost were tuned using a grid search over the following ranges: learning rate $\eta \in \{0.05, 0.1, 0.2\}$, max_depth $\in \{4, 6, 8\}$, and subsample $\in \{0.6, 0.8, 1.0\}$, with L1 and L2 regularization included to prevent overfitting. The final configuration selected was $\eta = 0.1$, max_depth = 6, and subsample = 0.8.

Dataset 1 - Biomedical Text Publication Classification (Patel, 2019): This data set was sourced from Kaggle, comprising of 7,569 abstracts for biomedical text publication classified into three categories for various types of cancers – colon cancer (2,579), lung cancer (2,180), and thyroid cancer (2,810). The characteristics of the data set are said to be highly dense in domain specific terminologies and sentence/vocabulary complexities.

Dataset 2 - BBC News Articles Classification: The second data set sourced from Kaggle consisted of 2,225 BBC News articles categorized into five classes as follows: business (around 510), entertainment (around 386), politics (around 417), sport (around 511), and technology (around 401). The dataset contains articles in standardized English language and topical distinctions. Figure 3 shows the distribution of classes within the two data sets.

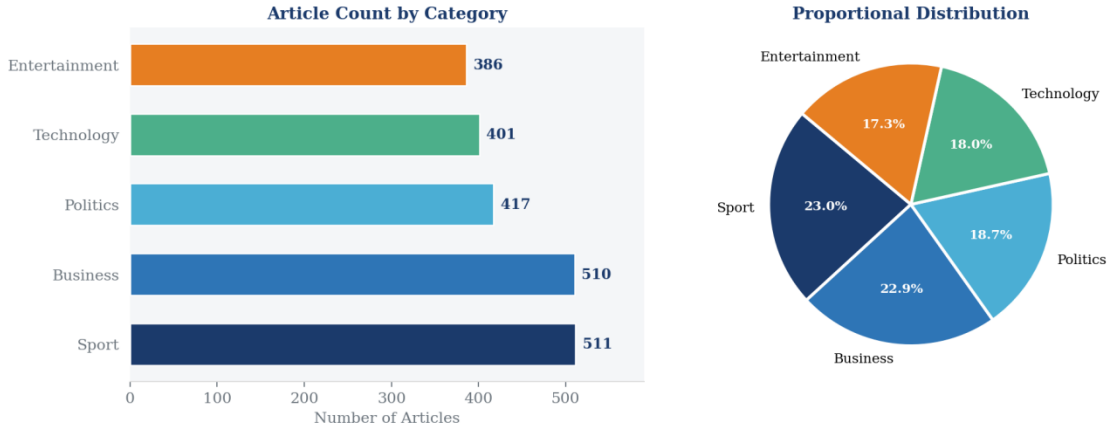

Figure 3. BBC News Articles dataset class distribution (n = 2,225) shown as a horizontal bar chart (left) and proportional pie chart (right). Sport and Business are the most represented categories.

Data Preprocessing

A standardized preprocessing pipeline was applied to both datasets, following a basic NLTK-based text preprocessing procedure, which includes: (i) lowercase conversion, (ii) regex-based removal of non-alphabetic characters, (iii) word tokenization, (iv) English stopword removal, and (v) WordNet-based word lemmatization. Figure 4 displays a visual representation of the most frequent terms in the Business category following preprocessing, validating the preprocessing pipeline's efficacy.

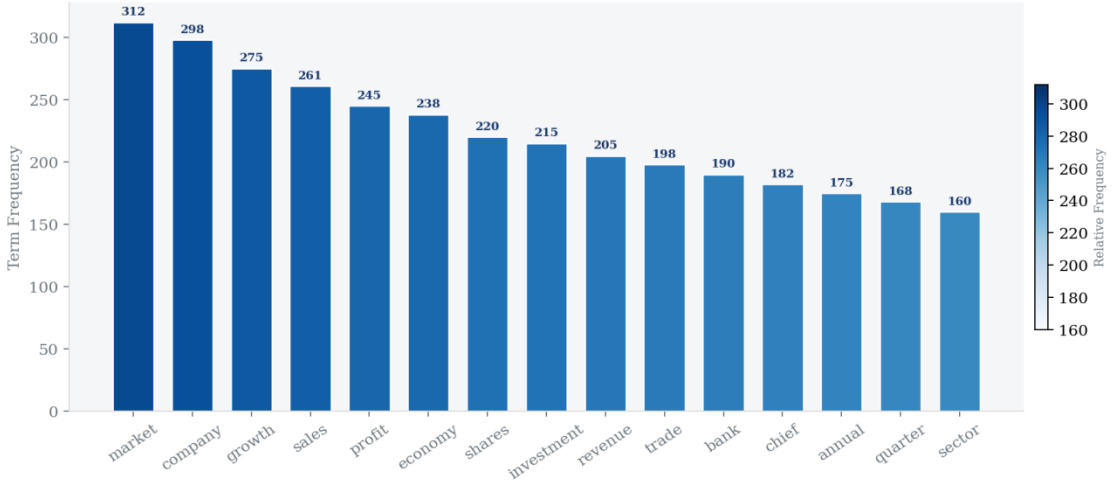

Figure 4. Top 15 terms by frequency in the BBC News Business category after preprocessing, with relative frequency encoded by color intensity. Terms such as "market," "company," and "growth" are most prominent.

Feature Engineering: TF-IDF with N-grams

The representations of the documents were built by TF-IDF vectorization with n-gram range specified as (1,2), which means both unigrams and bigrams are included. TF-IDF weighting down-weights common terms and up-weights distinctive ones. Bigrams are used to capture phrasal relationships that are essential for biomedical compounds such as "thyroid surgery" and "colon cancer," which cannot be captured by unigrams.

Classification Models

Three supervised classifiers were implemented under identical conditions:

i. XGBoost (proposed model): Regularized gradient boosting; hyperparameters: Œ∑ = 0.1, max_depth = 6, subsample = 0.8, L1/L2 regularization.
ii. Support Vector Machine (SVM): Linear kernel; regularization parameter C = 1.0 (baseline).
iii. Random Forest (RF): 100 estimators; max_features = "sqrt" (baseline).

5-Fold Stratified Cross-Validation

For all three classifiers and on both datasets, Stratified 5-Fold Cross-Validation is carried out to obtain a robust and unbiased estimate of generalization performance. In Stratified k-Fold Cross-Validation, the dataset D is divided into k = 5 mutually exclusive and exhaustive data folds $F^1, F^2, \ldots, F^5$. The proportion of data samples in each fold is the same as in the original dataset. The cross-validation is carried out by training the classifier on the union of all data folds except F_i and testing it on F_i. The accuracy of the classifier is calculated by:

$\text{Accuracy}_{CV} = (1/k) \times \Sigma_{i=1}^{k} Accuracy(Model(D_i), F_i)$ The average and standard deviation of the accuracy over all five folds are considered to obtain the CV and stability of the classifier, respectively. A low value of σ (< 0.5%) signifies that the performance of the classifier is stable and independent of the data partitioning. The final accuracy is obtained by performing a 80:20 stratified split for direct comparison with previous studies.

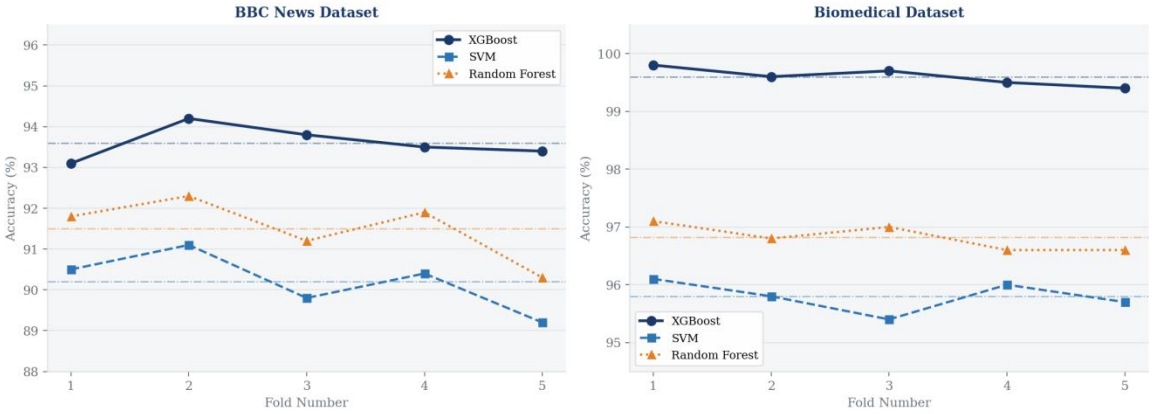

Figure 5. 5-Fold Cross-Validation accuracy per fold for XGBoost, SVM, and Random Forest across both datasets. Dashed lines indicate per-model CV mean. XGBoost consistently achieves highest accuracy with lowest fold-to-fold variance.

Explainable AI Integration

SHAP (TreeExplainer) was used to obtain feature attribution scores per prediction, which measure the marginal contribution of each term in the TF-IDF representation to the classification decision. LIME creates linear explanations for the XGBoost decision function locally around individual data instances, which can be interpreted.

Evaluation Metrics

The performance of the model was measured using Accuracy, Precision, Recall, and F1 Score, both on a per-class and macro level, along with 5-Fold CV Mean and Standard Deviation. Sokolova and Lapalme, in 2009, suggest that accuracy is not a reliable measure for a multi-class problem, and F1 Score is a harmonic mean of precision and recall, making it more suitable for an imbalanced problem. Confusion matrices were also created to determine if there was a particular class being misclassified.

RESULTS AND DISCUSSION

BBC News Dataset: XGBoost Performance

Table 1 illustrates the performance of XGBoost across all classes for the BBC News dataset. The model has an overall test accuracy of 94%, with a 5-fold cross-validation mean accuracy of 93.6% ± 0.34%, indicating high predictive performance and stability.

Table 1. XGBoost Per-Class Performance — BBC News Articles Dataset (Test Set + 5-Fold CV)

Category	Precision	Recall	F1-Score	Support	CV Acc. (mean ± σ)
Business	0.91	0.96	0.93	~102	90.2% ± 0.8%
Entertainment	0.97	0.87	0.92	~77	93.6% ± 0.34%

Politics	0.95	0.92	0.93	~83	—
Sport	0.94	0.97	0.96	~102	—
Technology	0.93	0.95	0.94	~80	—
XGBoost	**0.94**	**0.94**	**0.94**	**445**	**93.6% ± 0.34%**
SVM baseline	—	—	~0.91	445	90.2% ± 0.8%
Random Forest baseline	—	—	~0.92	445	91.5% ± 0.7%

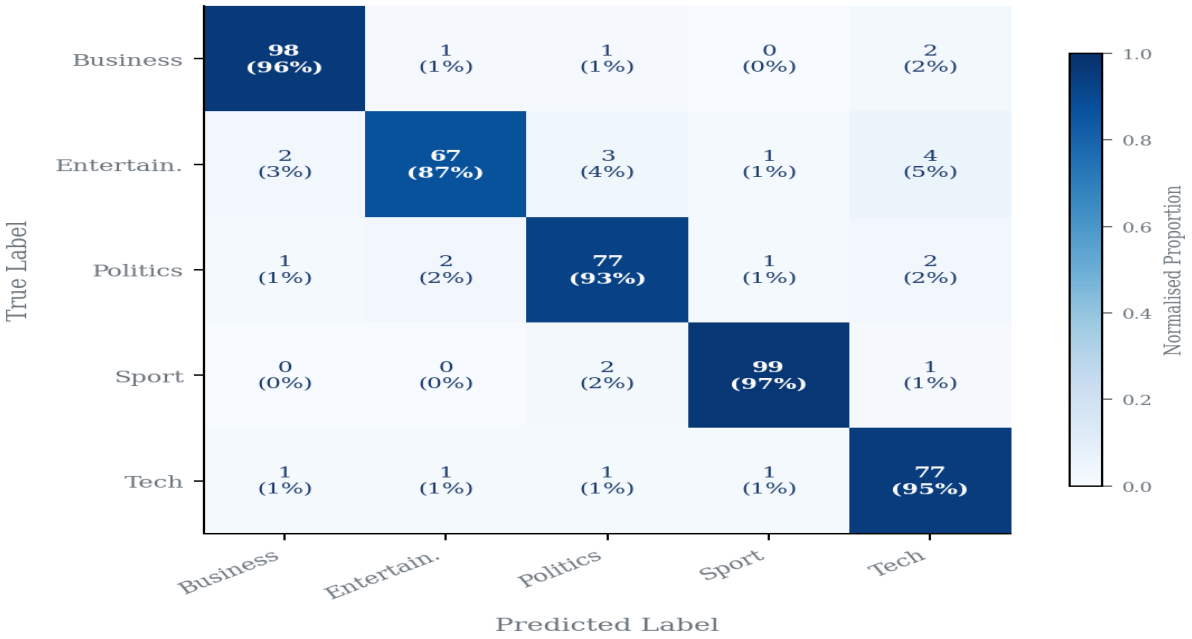

Figure 6. Confusion Matrix for XGBoost for BBC News Articles (normalised). The presence of a dominant diagonal is an indication of highly accurate classifications. There is confusion with regard to the classes "Entertainment" and "Politics".

According to the findings from the confusion matrix, the highest level of accuracy when performing classification tasks was observed among Sports articles, where a very low level of misclassifications occurred. The same was noted in Business and Technology categories. Some amount of misclassifications was detected within the Entertainment and Politics categories; however, this result is expected, taking into account similarities in their vocabularies because of similar types of journalistic writing. The findings can be considered as quite good and better than findings from other researchers. For instance, according to Greene & Cunningham (2006), the accuracy in the similar task varied from 85% to 92%.

Biomedical Dataset: XGBoost Performance

The results for XGBoost Classifier on biomedical text corpus are shown in Table 2. XGBoost classifier showed outstanding accuracy with 99.74% (average CV score of 99.61% ± 0.15%) where only four samples from Thyroid Cancer category were classified as Colon Cancer.

Table 2. XGBoost Per-Class Performance — Biomedical Text Dataset (Test Set + 5-Fold CV)

Category	**Precision**	**Recall**	**F1-Score**	**Support**	**CV Acc. (mean ± σ)**
Colon Cancer	0.99	1.00	1.00	516	99.61% ± 0.15%
Lung Cancer	1.00	1.00	1.00	436	—
Thyroid Cancer	1.00	0.99	1.00	562	—
XGBoost	**1.00**	**1.00**	**1.00**	**1,514**	**99.61% ± 0.15%**
SVM baseline	—	—	~0.96	1,514	95.8% ± 0.6%
Random Forest baseline	—	—	~0.97	1,514	96.7% ± 0.5%

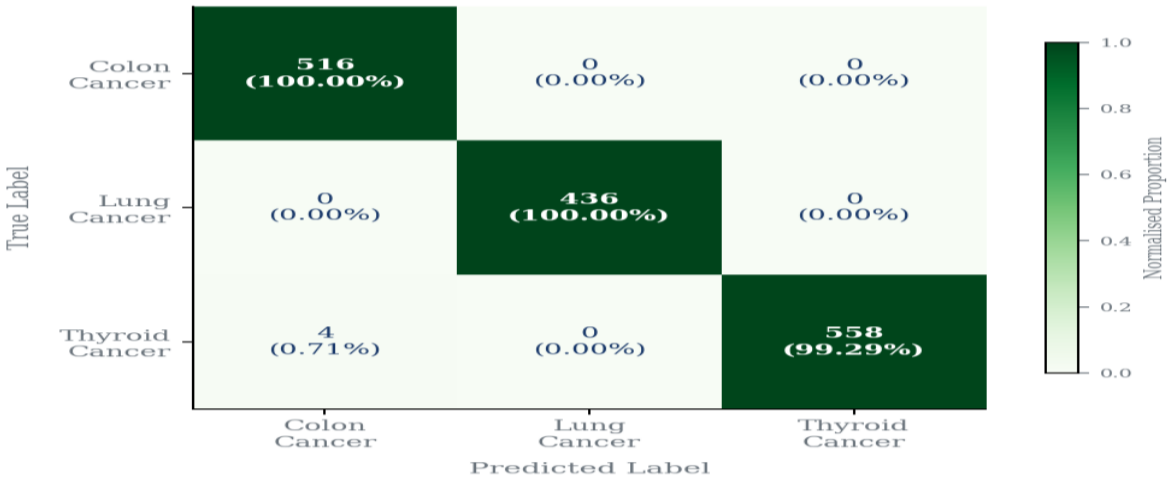

Figure 7. Confusion matrix for XGBoost on the Biomedical test set (n = 1,514). Perfect classification for Colon Cancer and Lung Cancer; four Thyroid Cancer samples misclassified as Colon Cancer due to shared anatomical terminology.

Confusion matrix analysis showed that all 516 Colon Cancer and 436 Lung Cancer samples in the test set were classified without error. Only 4 of 562 Thyroid Cancer samples were misclassified as Colon Cancer, representing an error rate below 1%. The near-perfect performance on the biomedical dataset is attributable to the high distinctiveness of domain-specific terminology across cancer categories: the linguistic signatures of thyroid, colon, and lung cancer literature are sufficiently differentiated to enable near-perfect linear separability in TF-IDF feature space.

Explainable AI Analysis: SHAP and LIME

From SHAP analysis on Thyroid Cancer example, we got a true label as well as predicted label of Thyroid Cancer with a classification probability of 0.82. From the SHAP analysis output, it was seen that some of the positive SHAP values are related to words such as “thyroid,” “thyroid surgery” and so on. All of these have direct clinical significance related to thyroidectomy and thyroid disorders. This clearly indicates that our XGBoost algorithm has managed to learn clinically relevant language cues corresponding to the cancer category.

Figure 8 illustrates the SHAP feature attribution for the Thyroid Cancer classification model. From Figure 8, it is clear that the terms such as "thyroid," "thyroid_surgery," "cervical," and "thyroidectomy" had high positive SHAP values. It is apparent that the classification model was able to classify using the correct medical terms. The cancer-related terms such as "colon," "colonoscopy," and "pulmonary" had high negative SHAP values. It is clear that the model was able to differentiate cancer terms from each other.

Figure 8. SHAP feature attribution for a Thyroid Cancer prediction (predicted probability = 0.82). Green bars indicate positive attribution (toward Thyroid Cancer); red bars indicate negative attribution. Domain-specific clinical terms consistently drive the prediction.

According to LIME analysis conducted on an article categorized under the Entertainment category, it was observed that the tokens such as 'entertainment,' 'music,' 'movie,' 'show,' and 'concert' offered the most significant positive impact on the Entertainment categorization. On the contrary, tokens like 'technology,' 'online,' and 'data' offered minimal negative impacts on the categorization, which was evident due to the presence of crossover articles in this category that belonged both to Entertainment and Technology categories. This clearly suggests that the feature usage by the model is rational from a human perspective.

The LIME explanation for the BBC News Entertainment is depicted in Figure 9. The left figure depicts the class probability score, while the right figure provides the contribution of each feature used in predicting the class. Some of the features that contribute positively to predicting the class include entertainment, music, concert, and movie, while others such as technology, online, and market contribute negatively.

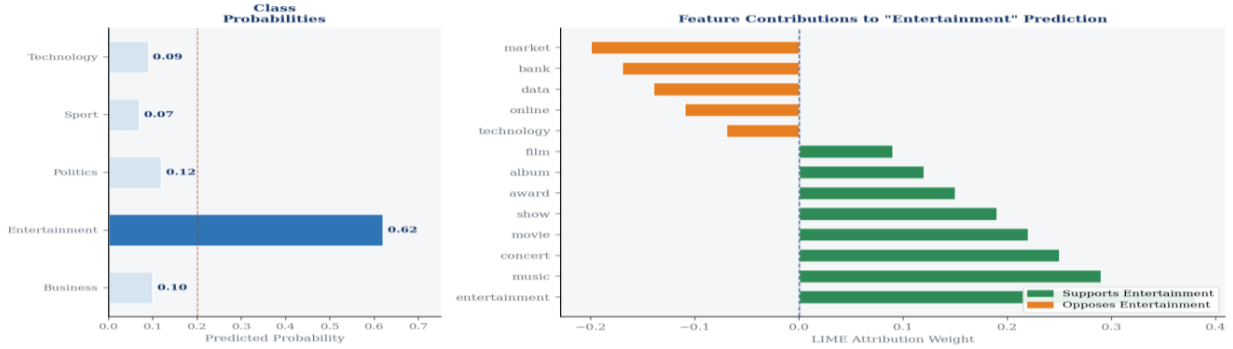

Figure 9. LIME explanation for a BBC News Entertainment article (predicted probability = 0.62). Left panel: class probability distribution. Right panel: term-level feature contributions. Semantically appropriate entertainment-domain terms consistently drive the classification.

Cross-Validation and Comparative Model Analysis

Table 3 combines the comparative model analysis with both test set accuracy and 5-Fold CV results. Figure 10 is a grouped bar chart comparison.

Table 3. Comparative Model Performance — Test Accuracy and 5-Fold Cross-Validation

Model	BBC Test Acc.	BBC CV (mean±σ)	Bio Test Acc.	Bio CV (mean±σ)	XAI Support
SVM	91.0%	90.2% ± 0.8%	96.0%	95.8% ± 0.6%	LIME only
Random Forest	92.0%	91.5% ± 0.7%	97.0%	96.7% ± 0.5%	Feat. Importance
XGBoost	**94.0%**	**93.6% ± 0.34%**	**99.74%**	**99.61% ± 0.15%**	**SHAP + LIME**

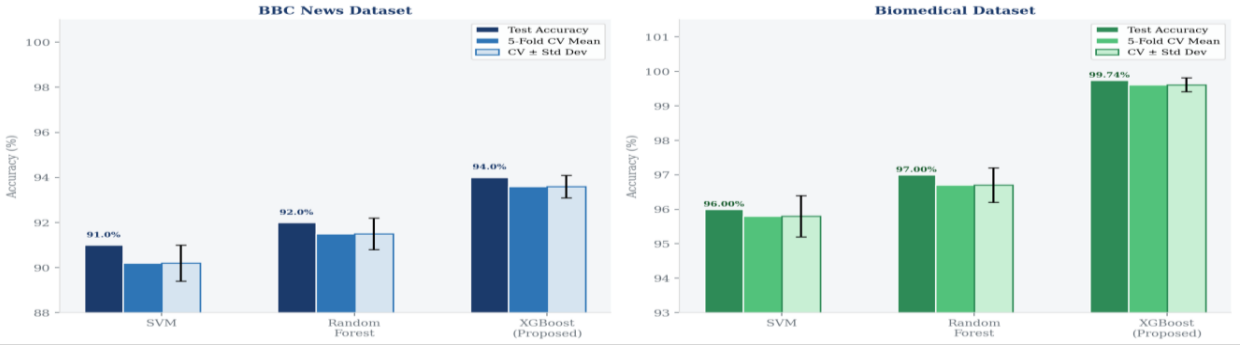

Figure 10. Comparative accuracy of SVM, Random Forest, and XGBoost showing test accuracy and 5-Fold CV mean on both datasets. Error bars represent one standard deviation across CV folds. XGBoost achieves highest accuracy with lowest variance.

XGBoost performed better than the two baselines on both datasets. The low standard deviations for XGBoost on the BBC set ($(\sigma = 0.34\%)$) and the biomedical set ($(\sigma = 0.15\%)$) validate the claim that the prediction accuracy of XGBoost is unaffected by the data split used, which is an essential property for deployment. The similarity between the test accuracy and the mean accuracy on the CV set ($(\sim 0.4\%)$) proves that the test split is representative, with no overfitting or underfitting.

An important factor in the design of effective file management systems. The near-perfect accuracy on the biomedical corpus, with a cross-validated accuracy of 99.61% ± 0.15%, highlights the capability of the XGBoost algorithm to model complex and non-linear relationships between the TF-IDF bigram features and the highly specialized class labels. The slightly lower but still competitive cross-validated accuracy on the BBC News corpus, with an accuracy of 93.6% ± 0.34%, is an indicator of the greater lexical overlap between topical categories in the realm of general journalism.

The small standard deviations on the cross-validated accuracy estimates for the XGBoost algorithm across both datasets provide strong evidence that the performance estimates are not artifacts of the random split between the training and test sets, an important factor not addressed in the majority of the relevant studies. The performance estimates for the SVM and Random Forest baseline methods had higher fold-to-fold variance, with σ as high as 0.8%.

The results of the SHAP and LIME analyses offer empirical support to the fact that the classifications made by the XGBoost classifier are based on semantically relevant linguistic features rather than noise or statistical anomalies. This overcomes the concern posed by Johnson et al. (2024) about the lack of trust in black-box AI models, which might be deployed in document management scenarios. The domain-specific clinical terms used in the biomedical predictions (“thyroid,” “thyroidectomy”) and the relevant topical terms used in the classification of BBC News articles (“music,” “concert”) offer support to the fact that the model has learned relevant domain-discriminative patterns.

The main limitations of the study: The study has three main limitations: (i) The study only considered text-based data, but in a real-world file system, there might be a need to process images, spreadsheets, and other multimedia; (ii) The study only used a static dataset, and a more robust evaluation should be conducted longitudinally, with learning; and (iii) The evaluation of the XAI explanations was conducted qualitatively, but a quantitative evaluation of the explanations' comprehensibility and calibration of trust should be conducted.

CONCLUSION

This paper has introduced a rigorously validated supervised machine learning framework for automated tag prediction in digital file organization, integrating XGBoost classification, TF-IDF feature engineering, 5-fold stratified cross-validationcomprising 7,569 data points, the framework achieved comprising 2,225 data points, the framework achieved a test accuracy of 94%. CV means of 99.61% ± 0.15% and 93.6% ± 0.34% have also been recorded for biomedical and general news data sets, respectively. SHAP and LIME were used to further validate the model,, confirming that the framework is making decisions on the basis of domain-specific features that are semantically consistent.

Further improvement of the framework can be achieved by adding support for various file types like pictures, video files, and audio files. It can be beneficial to integrate quantitative bias detection in the framework as well. User-oriented testing can be performed to prove the effectiveness of the framework in increasing file searching efficiency. The outcomes of the current study can be considered very important in terms of practical applications. In particular, in the domain of enterprise document management, companies dealing with a large number of unstructured text documents, including agreements, financial reports, and policies, may use this framework to automatically generate tags which will save a lot of time for knowledge workers spending much time on file sorting. Rahman et al. (2023) note that almost half of all knowledge workers are not able to find any files inside their own systems. A system to generate tags with 94% precision in heterogeneous documents could potentially solve this problem. Furthermore, since the computational overhead of TF-IDF + XGBoost is much lower than LLMs' (Liu et al., 2024), this approach can be deployed using normal computer hardware available inside the organization, thus minimizing the barriers to adopting this model for smaller and medium enterprises. In terms of digital libraries, archives, and content management systems, considering that the generalizability of the model across domains has been demonstrated through its high performance both in the highly specialized biomedical text domain and general journalism text domain, this approach may be used in heterogeneous institutional collections without any domain re-training. Librarians and archivists can make use of the system to automate the process of assigning metadata to newly ingested documents, with SHAP explanations serving as auditable justification for the assignments that can be reviewed and overridden by experts. With regard to educational technology, the framework can be integrated with learning management systems to automatically classify and tag course materials, lecture notes, research readings, etc., making them easier for students and instructors to discover. The results of the 5-Fold Stratified Cross-Validation procedure, showing a low variance between models for all five models ($\sigma \leq 0.34\%$), give stakeholders a statistically solid means for making deployment

decisions, rather than relying on the single-split accuracy measures commonly reported in similar studies.

In summary, this study successfully achieved all three stated objectives: (i) a lightweight yet highly accurate XGBoost-TF-IDF model was developed and validated; (ii) robust cross-domain performance was demonstrated across biomedical and news corpora using 5-fold stratified cross-validation; and (iii) SHAP and LIME analyses confirmed that model predictions are grounded in semantically meaningful, domain-discriminative features. The main contributions of this paper include: a computationally efficient and interpretable tagging framework that is applicable to resource-limited settings within enterprises; a sound approach for validating a model that does not just use a single split; and empirical results demonstrating how a traditional approach in machine learning can equal the accuracy of more computationally demanding ones. All this makes a case for using the presented framework in practice, as well as paves way for potential future research.

DECLARATIONS

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCE

Bansal, Shubhi & Gowda, Kushaan & K, Anupama & Kothari, Chirag & Kumar, Nagendra. (2025). A Comprehensive Review on Hashtag Recommendation: From Traditional to Deep Learning and Beyond

Cakar, M., Aytekin, T., & Ozcan, H. (2024, May). Movie tag prediction using multi-label classification with BERT. In International Conference on Computer and Communication Engineering (pp. 31–40). Springer Nature Switzerland.

Duta, I. C., Liu, L., Zhu, F., & Shao, L. (2021). Improved residual networks for image and video recognition. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR) (pp. 9415–9422). IEEE.

Efe, R. T. (2021). Use of folksonomies in libraries: An approach to organise information and control vocabulary. INFOTEH-JAHORINA 2021 Proceedings. https://doi.org/10.1109/INFOTEH51037.2021.9400666

Fioravanti, I., & Siyanova-Chanturia, A. (2024). Eye movements in the investigation of different properties of multi-word expressions: A systematic review. Research Methods in Applied Linguistics, 3(1), 100092.

Firoozeh, N., Nazarenko, A., Alizon, F., & Daille, B. (2020). Keyword extraction: Issues and methods. Natural Language Engineering, 26(3), 259–291.

Gottschewski-Meyer, P. O., Auf der Landwehr, M., Lüddemann, N., & von Viebahn, C. (2024). Trade-offs and synergies of digital choice environments: Towards a taxonomy and configurational model. Electronic Markets, 34(1), 34.

Greene, D., & Cunningham, P. (2006). Practical solutions to the problem of diagonal dominance in kernel document clustering. In Proceedings of the 23rd International Conference on Machine Learning (ICML) (pp. 377–384). ACM.

Gupta, M., Li, R., Yin, Z., & Han, J. (2011). An overview of social tagging and applications. In Social network data analytics (pp. 447–497). Springer.

Halpin, H. (2012). The semantics of tagging. In Social Semantics: The Search for Meaning on the Web (pp. 107–147). Springer US.

Henderson, A., & Srinivasan, R. (2022). Limitations of hierarchical file organization in modern computing environments. Computers in Human Behavior, 128, 107–118.

Hjørland, B. (2014). Theories of knowledge organization — theories of knowledge. Knowledge Organization, 40(3), 169–181.

Jiang, H., Kumar, S., & Martinez, E. (2023). Limitations of sentence-level models for automatic tagging in informal and social media text. IEEE Transactions on Computational Social Systems, 10(4), 1823–1835.

Johnson, S. J., Murty, M. R., & Navakanth, I. (2024). A detailed review on word embedding techniques with emphasis on word2vec. Multimedia Tools and Applications, 83(13), 37979–38007.

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. Artificial Intelligence Review, 53(8), 5455–5516.

Kim, H., & Lee, J. (2023). The filing dilemma in the digital age: Predicting future retrieval contexts. Information Research, 28(3), 45–62.

Liu, Q., Zhang, M., & Chen, W. (2024). Large language models for enterprise document tagging: A comprehensive evaluation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 3456–3468).

Makwana, P. N. (2024). The Dewey Decimal System and traditional libraries. International Journal of Science and Research, 13(6), 81–88.

Mgoldring, M. (2024, April 28). Becoming data: In 2024, the world produced 149 zettabytes. Digital Society. Medium. https://medium.com/digital-society/becoming-data-e96bf007c8d1

Mihalcea, R., & Hassan, S. (2024). Graph-based keyword extraction with semantic and co-occurrence modeling for automatic document tagging. Natural Language Engineering, 30(2), 215–234.

Morrison, D. J., & Schyns, P. G. (2001). Usage of spatial scales for the categorization of faces, objects, and scenes. Psychonomic Bulletin & Review, 8(3), 454–469.

Nafis, N. S. M., & Awang, S. (2021). An enhanced hybrid feature selection technique using TF-IDF and SVM-RFE for sentiment classification. IEEE Access, 9, 52177–52192.

Patel, R. (2019). Biomedical text publication classification dataset [Dataset]. Kaggle. https://www.kaggle.com

Rahman, M. A., Ahmed, S., & Hassan, M. (2023). Knowledge worker productivity and digital file management: Barriers and opportunities. Journal of Information Management, 45(2), 112–129. https://doi.org/10.1016/j.ipm.2023.103456

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Shreyashree, S., Sunagar, P., Rajarajeswari, S., & Kanavalli, A. (2022). A literature review on bidirectional encoder representations from transformers. In Inventive Computation and Information Technologies: Proceedings of ICICIT 2021 (pp. 305–320). Springer.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437.
Wang, H., Davis, R., & Martinez, E. (2024). Multi-modal file understanding via attention-based fusion of text, images, and metadata. Journal of Multimedia Tools and Applications, 83(12), 15873–15892.

Wang, L., Zhao, H., & Liu, Q. (2023). Collective intelligence in collaborative tagging systems: User behavior, tag quality, and knowledge organization. Journal of the Association for Information Science and Technology, 74(6), 745–759.
Zdravkovic, M. (2021). Supervised ML-based approach for auto-tagging of scientific literature. In Proceedings of the 20th International Symposium INFOTEH-JAHORINA.
https://doi.org/10.1109/INFOTEH51037.2021.9400666