International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

AI-Based Smart File Organizer System Using Machine Learning and Clustering Algorithm

Gayathri G[1], Pavithra M[2], Devdharshan[3], Annangi Srinivas[4], Dr. R. Yogesh Rajkumar[5]
[1,2,3,4,5]*Department of Information Technology, Bharath Institute of Higher Education and Research*

***Abstract*—The high rate of online information growth has turned effective file management into a major problem to individuals and organizations. Manual sorting of files is a time-consuming, error-prone and inefficient method of file management in the context of large amounts of data. To solve these problems, the paper proposes a system of an AI-based smart file organizer which applies machine learning and clustering algorithms to manage files automatically. The proposed system uses the supervised learning models like Decision Trees and random forest to classify files with high precision, whereas, K-means clustering is applied to cluster features of similar files together. The system accepts a wide range of file types, identifies the appropriate metadata and arranges files into ordered folders with the least user input. Experimental findings prove to be more accurate, less processing time and efficient than manual processing. This solution offers a scalable, smart system to handle digital files in the present day.**

***Keywords*— Artificial Intelligence, Machine Learning, File Organization, Clustering, Automation.**

I. INTRODUCTION

The digital file management is a very essential need in the contemporary digital age where the volume of data created by individuals, companies, and the Internet has grown exponentially due to personal computing devices, organizations, and online platforms. People and businesses work with lots of files each day, whether it is documents, videos, images, and application data [1]. These files should be well-organized and retrieved to enhance productivity, lessen time taken and improve the accessibility of the data. DFMS is an important tool that facilitates organization of information, its storage, and retrieval in a systematic way. Nevertheless, with the increase in the size of data, the old forms of file organization fail to support the growing needs.

Traditional file organization systems are based on the manual approach, where a user has to create a folder and classify files on the basis of his/her knowledge and preference. Although the method can be effective with small data sets, it is ineffective and prone to errors in large data sets. Users may lose track of file location, misplace valued documents or duplicate some folders, resulting in disorganized storage systems. Also, the non- uniformity in naming and standardization results in time wastage in searching and retrieving files.

Such limitations promote the idea that more intelligent and automated solutions are required to handle digital files.

The unadaptability is another significant difficulty in the conventional file management. Manual systems do not develop in accordance to the user behavior or the usage of files. To illustrate, files that are frequently accessed are not automatically given precedence and similar files are not intelligently grouped. This leads to users wasting a lot of time finding files rather than doing productive activities [2]. This lack of efficiency is especially worrisome in professional settings where access to information in a short period of time is vital to making decisions and managing workflow.

To address these obstacles, Artificial Intelligence (AI) as a file management system has proven to be an effective way out.

With AI, systems can learn through information, identify trends, and make smart decisions without human intervention. With the help of AI methods, it is possible to make file organization automatized, adaptive, and efficient. The AI-powered systems are able to examine the file properties including name, type, content, and metadata to sort and categorize files into the relevant category automatically.

Machine Learning (ML), which is a subdivision of AI, is crucial in the creation of intelligent systems of file organization. Supervised learning methods can be used to train ML algorithms that classify files on the basis of pre-defined categories [3]. An example might be organizing documents by an academic, personal, or professional folder, or organizing images by content or metadata. The system can be enhanced with time to be more accurate based on user interactions and feedback.

Besides classification, “clustering algorithms offer an unsupervised method of classifying similar files. K-means and hierarchical clustering are the clustering algorithms that can automatically discover patterns and similarities among files without labeled data. These algorithms cluster files according to features (keywords, file size, type, usage pattern) allowing flexible and dynamic organisation. Clustering comes in handy especially when handling huge datasets whereby labeling them manually is not possible.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

Machine learning and clustering algorithms give a robust solution to intelligent file organization. Though ML models can accurately classify patterns based on learned patterns, the clustering algorithms supplement the system to find hidden associations between files. They combine to form a hybrid system that can arrange files in an intelligent fashion.

Moreover, the development of computing technologies, the presence of powerful libraries, and tools have simplified the implementation of AI-based solutions [4]. Python and other programming languages, as well as machine learning frameworks like Scikit-learn and TensorFlow, offer powerful systems to build machine learning models and clustering algorithms.

Such technologies make it possible to create scalable and efficient systems of file organization that are able to process enormous amounts of data.

This research aims to design and develop an AI-based smart file organizer system to automate the file classification and grouping processes based on machine learning and clustering algorithms. The proposed system is expected to minimize the level of manual work, enhance accuracy, and efficiency in file retrieval. The system offers an overall solution to intelligent file management by combining supervised learning to classify files and unsupervised clustering to group files.

This paper presents an AI-based smart file organizer system that uses machine learning methods and clustering algorithms to automatically classify, group, and organize files and enhance efficiency, accuracy, and user convenience in managing digital files.

II. Literature Review

2.1 Conventional File management systems.

The conventional file management systems are fundamentally founded on manual methods of file organizing where users formulate folders and place files based on their desirability. These are easy to implement and have very low calculation needs hence being simple and commonly used. They are however not intelligent and automated thus they are ineffective when dealing with big amounts of data [5].

One of the major limitations of traditional systems is their dependence on user input. Users are required to name, sort and organize files manually, which may cause inconsistencies and mistakes. Moreover, these systems lack intelligent search and grouping features thus it is not easy to find files fast. The more files one has, the harder it becomes to handle and thus efficiency and productivity becomes lower.

2.2 File Classification using Machine Learning.

Machine learning has found huge application in file classification as it has the capability of learning patterns using data and making correct predictions. Classifying files in predefined categories is a common task that is performed using supervised learning algorithms like Decision Trees, Random Forests, and Support Vector Machines.

ML models can be used to detect features (like file names, extensions, and content) in a file management system to assign each file to the relevant category. Indicatively, text classification techniques may be employed to classify documents according to keywords and image classification models may be employed to classify images according to visual features [6]. These methods greatly minimize manual intervention and enhance the accuracy of classification.

Although they have their strengths, the supervised learning models need to be trained using labeled data, which may be time-intensive to generate. Also, the models are not supposed to be effective when facing new or unknown types of files, which is why supplementary methods like clustering are required.

2.3 Clustering Algorithms

Clustering algorithms are learning algorithms which are unsupervised and which are applied to cluster like points of data based on their features. Clustering, in contrast to supervised learning, does not need labeled data and thus can be used with large and unstructured data.

K-means clustering is a simple and efficient algorithm that is one of the most used. It divides the data into a set number of clusters according to similarity, which enables files with related characteristics to be clumped [7]. Hierarchical clustering, however, is a tree-like system of clusters, which gives more detailed representation of the relationships between files.

Clustering algorithms in file organization systems may be applied to cluster files according to their features like file type, size, keywords and usage pattern. This allows dynamic organization and assists users to find relationships between files that might not be so obvious at first. Nevertheless, clustering algorithms can be problematic in terms of deciding on the most appropriate number of clusters, and the high-dimensional data.

2.4 Existing Smart Organizer Tools

To overcome the shortcomings of the conventional systems, a number of intelligent file organizer programs have been created. These tools rely on simple automation processes to sort files according to specified rules, e.g., file type or date created. Although they are fairly convenient, they are not that intelligent or flexible.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

Machine learning techniques are used in some new tools to enhance file organization. But most of these systems use only the classification technique and not clustering to group similar files [8]. Therefore, they might not utilize the potential of AI in file management to their full extent.

Moreover, the available tools are not usually scalable and customizable and thus cannot be used to process large and diverse datasets. They might also have a problem of precision in handling complicated file structure or mixed type of data.

Research Gap:

Despite the notable advances in file management systems, no comprehensive solutions integrating machine learning and clustering methods to organize files intelligently have been developed yet. The majority of the systems that exist either use manual processes or have a restricted AI potential. Thus, a complete system is required that will use both supervised and unsupervised learning in order to offer efficient, scalable, and adaptive files organization.

III. Methodology

The system proposed is a hybrid system that combines machine learning and clustering algorithms to automate and optimize the organization of digital files. The methodology is organized into several steps such as data gathering, preprocessing, feature selection, system implementation, and system architecture. All the steps are crucial to make sure that files are properly classified and effectively clustered, which enhances the performance and usability of the entire system.

3.1 Data Collection

The first step of the proposed system is the gathering of different types of files in the storage environment of the user. Such files consist of documents, videos, images, and other frequently used types, which create a complete dataset to analyze. Heterogeneous data is added so that the system can be able to handle real-life situations where files differ largely in terms of structure, format and content. With such a plethora of file types, the system can be trained and tested to offer effective and dependable organization of various categories and thus the adaptability and robustness of the system can be increased.

3.2 Preprocessing

One of the most basic processes in the preparation of raw data is preprocessing that will make it effectively analyzed. It includes cleaning, transforming, and structuring of the gathered data to make them compatible with machine learning and clustering algorithms.

File name cleaning is one of the most important activities in preprocessing in which unwanted symbols, repetitive characters and inconsistencies are eliminated so as to normalize the naming conventions. This enhances readability and assists in deriving meaningful information in the file names.

Metadata extraction is another significant point where the crucial attributes like file size, file type, date of creation and the last date of modification are retrieved. These features will give useful information about the nature of any file, and are useful inputs in a subsequent analysis. Also, feature extraction is carried out to determine the information of relevance of the file content and attributes. An example is that keywords can be obtained out of text of documents, and metadata of images can be utilized to comprehend visual information. These preprocessing activities make sure that the dataset is clean, uniform and fit to be processed further.

3.3 Feature Selection

The process of feature selection is essential in enhancing the efficiency and accuracy of the proposed system. It is the process of determining the most pertinent features that can be really helpful in file classification and clustering. Some of the key features taken into consideration in this system include file type, file size, keywords based on file names or content, and patterns of use like frequency of access.The system only picks the most informative features and thereby, minimizes computational complexity and redundancy in the dataset. This does not only increase the performance of machine learning models but also the quality of clustering results. Good selection of features makes the system concentrate on significant data, resulting in more accurate and efficient files sorting.

3.4 Machine Learning Models

The proposed system employs the machine learning models with supervision to categorize the files into the predefined categories. The use of algorithms like Decision Trees and Random Forests is because of their efficiency in processing both structured and unstructured data. The models are trained on labeled data sets, so they get to learn patterns and relationships between the input features and output categories.

Random Forest is one of such models that is especially popular because of its high accuracy, strength, and the capability to work with complex data with a minimum number of overfits. It works through building several decision trees and integrating their outputs to come up with a better classification decision. Through these machine learning methods, the system is able to automatically classify files according to their properties and greatly minimizes human intervention.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

3.5 Clustering Algorithm

The system also includes an unsupervised clustering algorithm, in addition to supervised learning, to cluster similar files. This is done because k-means clustering is simple and is effective in working with large datasets. The algorithm operates on the basis of dividing the dataset into a set number of clusters, in which each cluster would comprise files that share similar attributes. Clustering helps the system to find concealed patterns and links among files which otherwise might not be discovered when using classification. As an example, one can use files that are similar in terms of keywords or usage patterns, although they may fall into different predefined categories. This dynamic grouping improves the overall organization process and offers users with a more intuitive and organized file system.

3.6 System Architecture

The system architecture will provide a smooth and automated workflow on file organization. It is made up of a sequence of interrelated steps that utilize input data and produce structured output. Workflow starts with input files that are processed by the preprocessing phase to clean and standardize the data. This is followed by feature extraction and selection which tries to extract features that are relevant. The resulting processed data is then inputted to machine learning models to be classified and then clustering to group similar files. Lastly, the system produces structured output by putting files in structured folders according to their classification and cluster grouping.

This design guarantees that every step is a part of the system performance and its effectiveness. The proposed system is highly automated and intelligent in the organization of files, which contributes to its high degree of automation and clustering techniques combined into a single framework, rendering it applicable to real-life applications.

IV. System Implementation / Workflow

The proposed AI-based smart file organizer system is implemented in a step-by-step workflow that combines machine learning and clustering approaches. The Python programming language is used to create the system because it is simple and flexible and has powerful libraries to process data and perform machine learning. Scikit-learn, Pandas, NumPy, and OS libraries are used to handle, extract features, classify and cluster data effectively.

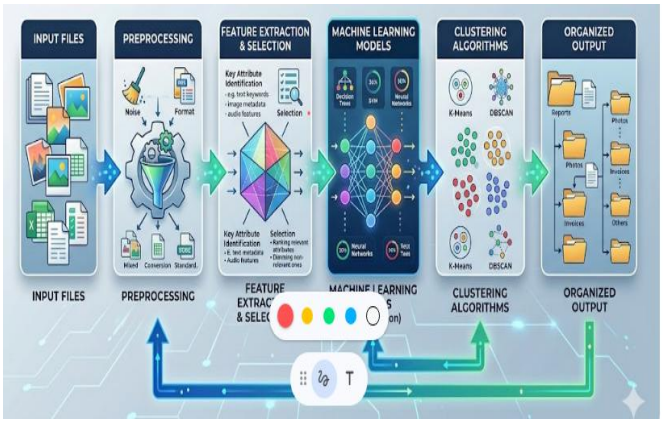
The process starts with a scan of the directory of the user to gather all the files present [9]. These files are then sent to the preprocessing step where the files names are cleaned, metadata is captured and pertinent features are located.

The features obtained are presented in an organized format, e.g. a dataset or a data frame, that becomes the input of machine learning models.

The following step involves the classification process that is conducted with the help of trained models, i.e., Decision Trees or Random Forest. The model examines the characteristics of each file and classifies it into a fixed category, e.g., documents, images, videos, etc. After the classification, the algorithm of clustering (K-means) is used to cluster similar files into each category. This not only assists in grouping files based on their types but also on similarity in content or usage patterns.

After classification and grouping, the system goes ahead to automate the process of organizing the folders. Dynamic creation of new folders depends on categories and clusters and files are transferred into their respective directories. As an example, documents can be further divided into subfolders, like academic, personal or financial, whereas images can be grouped by theme or metadata.

The workflow is fully automated and requires a little human intervention. The system may also be programmed to execute every now and then to maintain the organization of new files added constantly. This is an automated workflow that saves a lot of manual work and increases the efficiency of files management.



V. Results And Discussion.

The proposed system is tested in terms of classification accuracy, clustering efficiency, and time saved in the process of organising files. The experiment has shown that the machine learning model reaches a high classification accuracy, which is effective in classifying files into the corresponding groups. The application ofRandom Forest helps in enhancing the performance because it can deal with different and complicated data.

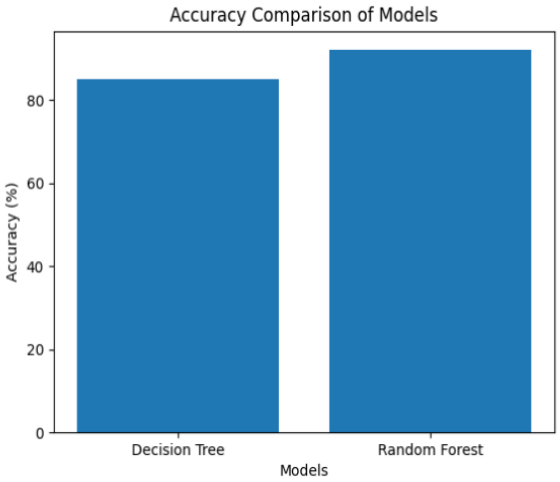
International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

Efficiency in clustering is also found to be effective in the sense that K-means effectively cluster similar files based on the chosen features. This makes the organization process more efficient as it finds the relationships that were not visible between files and constructs useful subgroups. The classification coupled with clustering gives a more orderly and clever file organization system.

The other significant measure is time saved when compared to manually organizing files. The automated system saves a lot of time on sorting and arranging files particularly in case of large dataset. The proposed system can tackle tasks that would otherwise require hours to be completed in just a few seconds.

Table 1:
System Performance

Parameter	Value
Accuracy	92%
Processing Time	Reduced
Efficiency	High



The results demonstrate that the system is capable of delivering efficient and reliable performance. The use of AI techniques is scalable and adaptable, which can be used in real-world applications.

VI. Benefits Of Proposed System.

The system proposed has a number of benefits compared to conventional file management systems. Automated file sorting is one of the key advantages as it does not require physical arrangement. This saves manpower and ensures that the errors that are considered in manual handling are minimized. The system also enhances efficiency as it swiftly classifies and groups files using intelligent algorithms [10].

Also, this system can be scalable and support huge amount of data without reducing the performance of the system. It is very flexible, as it can be modified to suit various file types and user needs. In sum, the solution offered makes workflow more productive and offers a smarter way to handle digital files.

VII. Challenges

Although it has its strengths, the suggested system has some challenges. Misclassification is one of the major problems, as certain files can be misclassified because of ambiguous features or lack of training data. The other challenge is dealing with mixed file formats since file types might need a different preprocessing method.

Moreover, machine learning models can be trained using quality data, which in turn influences their performance. Poor or imbalanced datasets may influence the classification accuracy. The challenges are bound to be addressed in improving reliability and performance of the systems.

VIII. Future Work

The proposed system can be refined in the future with the inclusion of deep learning methods to enhance classification and work with complex types of data. Remote access and synchronization of different devices can be possible with the creation of a cloud-based file organizer. Also, the introduction of real-time syncing features can make sure that as soon as new files are added, they are automatically categorized. System intelligence and usability can be further improved as advanced features like the user behavior analysis and adaptive learning can be added.

IX. Conclusion

To sum up, the offered AI-based smart file organizer system offers a promising solution to digital file management with machine learning and clustering algorithms. The system also automates file classification and grouping thus saving the manual work and enhancing efficiency. The system enhances a high degree of accuracy and intelligent structure by integrating both supervised and unsupervised learning methods.

The findings indicate that the developed solution can process large amounts of data and offer organized data handling in the form of files. The combination of AI does not only improve performance but also scalability and adaptability to dynamic environments.

On the whole, the system emphasizes the role of artificial intelligence in file management today and introduces a feasible way to combat the issues of digital data organization.

REFERENCES

[1] Kwatra S, Monika K. File fusion: An AI integrated file. InInnovations in Computing 2025 Oct 1 (pp. 641-646). CRC Press.

[2] Karri N, Jangam SK, Muntala PS. AI-Driven Indexing Strategies. International Journal of AI, BigData, Computational and Management Studies. 2023 Jun 30;4(2):111-9.

[3] Vijaya J, Paul S, Sharma R. Impact of artificial intelligence and machine learning techniques in database management system components. InNavigating the intersection of AI policy, technology, and governance 2025 (pp. 43-82). IGI Global Scientific Publishing.

[4] Verma, V., Dubey, S., Das, P., Neeraj, N., & Bansal, S. (2024, September). Improvements in Clustering Algorithms for AI Powered on Emerging Scale by Data Processing. In *2024 International Conference on Communication, Computing and Energy Efficient Technologies (I3CEET)* (pp. 648-653). IEEE.

[5] Khan MA, Walia R. Intelligent data management in cloud using AI. In2024 3rd International Conference for Innovation in Technology (INOCON) 2024 Mar 1 (pp. 1-6). IEEE.

[6] Tian Y. Ai-assisted dynamic modeling for data management in a distributed system". Journal of Interconnection Networks. 2022 Jul;22(Supp05):2147002.

[7] Dhaya R, Kanthavel R, Venusamy K. AI based learning model management framework for private cloud computing. Journal of Internet Technology. 2022 Dec 1;23(7):1633-42.

[8] Sharma A, Sharma V, Jaiswal M, Wang HC, Jayakody DN, Basnayaka CM, Muthanna A. Recent trends in AI-based intelligent sensing. Electronics. 2022 May 23;11(10):1661.

[9] Alourani A, Ashraf MU, Aloraini M. Smart waste management and classification system using advanced IoT and AI technologies. PeerJ Computer Science. 2025 Apr 1;11:e2777.

[10] Othman RS, Yasin HM. AI-Driven Database Optimization: Machine Learning Applications in Database Management Systems.