iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54035

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Decision Tree Learning Based Feature Selection and Evaluation for Image Classification

Saikiran Kutikuppala[1], Pranavi Kondamadugula[2], Pavan Sai Katakam[3], Pranay Reddy Kotla[4], Sai Sawanth Kanjarla[5], Yogesh Kakde (Guide)[6]
[1, 2, 3, 4, 5]*Student(s) - CSE - AI & ML, Malla Reddy University, Hyderabad*
[6]*Assistant Professor, CSE - AI & ML, Malla Reddy University, Hyderabad*

Abstract: ***The problem statement focuses on feature evaluation and selection for image classification using decision tree learning. The objective is to identify the most significant features in an image dataset and train a decision tree classifier using these selected features. The accuracy of an image classifier heavily relies on the quality and relevance of the features used to represent the images. Hence, it is crucial to identify the most important features and eliminate the irrelevant ones to enhance the classifier's accuracy. To implement this approach, we can utilize scikit-learn, a popular machine learning library in Python. The solution must involve training a decision tree classifier on the dataset and extracting feature importances, selecting the top features using modules from sklearn like "SelectFromModel", and also performing hyperparameter tuning using "GridSearchCV" and training a new decision tree classifier on the selected features with the best hyperparameters. Decision trees are a popular machine learning algorithm that uses a tree-like model of decisions and their possible consequences. By training a decision tree classifier on an image dataset and extracting feature importances, it is possible to identify the most important features and select them for use in a new decision tree classifier that can improve classification accuracy. It is important to note that decision tree learning is a versatile machine learning algorithm that can handle both binary and multi-class classification problems. Additionally, it is advantageous for feature evaluation and selection in image classification tasks. By identifying the most relevant features, this approach can enhance the accuracy of the classifier and reduce computational complexity, making it suitable for large datasets. By following this outlined approach, you can create a project that addresses feature evaluation, selection, and classification accuracy improvement using decision tree learning in the context of image classification.***

Keywords: feature evaluation, feature selection, image classification, decision tree classifier, accuracy improvement, feature importances, SelectFromModel, hyperparameter tuning, GridSearchCV, binary and multi-class classification, scikit-learn, classifier accuracy

I. INTRODUCTION

Image classification plays a crucial role in computer vision, finding applications in various domains like medical imaging, remote sensing, surveillance, and self-driving cars. The accuracy and effectiveness of an image classifier greatly depends on the features used to represent the images.

However, not all features are equally significant (contribute equally to the classification task), and some may even be irrelevant or redundant. Therefore, it becomes crucial to identify the most important features and eliminate the irrelevant ones to enhance the accuracy and efficiency of the image classifier. Decision tree learning provides an effective approach to feature evaluation and selection in image classification. Decision trees are widely used in machine learning due to their ability to model (outcomes) decisions and employ a tree-like structure to model decisions and their consequences. By training a decision tree classifier on an image dataset and assessing feature importances, it becomes possible to determine the most influential features. These important features can then be selected for use in a new decision tree classifier, leading to improved and enhanced classification accuracy.

The problem addressed in this project is feature evaluation and selection for image classification using decision tree learning. The primary objective is to develop an approach that utilizes decision trees to assess the quality and relevance of features, enabling effective feature selection. The project aims to demonstrate the effectiveness and scalability of this decision tree-based approach by applying it to diverse types of image data.

The outcome of the project will be a robust and effective decision tree-based approach that can be used to improve the performance and accuracy of image classification models.

II. RELATED WORK

An easy way to comply with IJRASET paper formatting requirements is to use this document as a template and simply type your text into it. The primary focus of this project is to address feature evaluation and selection for image classification using decision tree learning. The main objective is to enhance the accuracy of image classification models by identifying the most pertinent features and eliminating irrelevant or redundant ones. By reducing the dimensionality of the feature space and improving.
Related Work on decision tree-based feature selection and feature extraction for image classification:

A. Review on Feature Selection Techniques (Ref. Fig – 1 & 2)

1) *Filter-based Techniques:* Feature selection techniques and methods are essential components of machine learning and data analysis, serving the purpose of identifying the most informative and relevant features for a given task. The vast array of feature selection approaches developed over the years aims to tackle the challenge of dimensionality reduction and enhance the performance of predictive models. Among the commonly utilized methods is the filter-based feature selection, which employs statistical measures such as correlation, mutual information, or chi-square tests to rank features. While filter methods are computationally efficient and provide an initial assessment of feature importance, they often overlook the intricate relationships between features and the target variable, thus limiting their effectiveness.
2) *Wrapper-based Feature Selection:* In contrast to filter-based techniques, wrapper-based feature selection methods take into account the predictive performance of models when selecting features. These methods involve a search strategy, often implemented through brute force or heuristic algorithms, to evaluate various subsets of features by training and testing models on each subset. By incorporating the learning algorithm directly into the feature selection process, wrapper methods can capture the interactive effects between features and improve the accuracy of the resulting models. However, due to the high dimensionality of the search space, wrapper methods can be computationally expensive and prone to overfitting or instability issues.
3) *Embedded Feature Selection Methods:* These address the feature selection challenge by integrating the process within the learning algorithm itself. These techniques optimize the feature subset during the model training phase, leveraging the inherent feature importance estimation capabilities of the learning algorithm. Example of embedded methods include Lasso regularization, decision trees, and genetic algorithms. Embedded methods are particularly effective in handling high-dimensional data and automatically selecting relevant features, thereby mitigating the risk of overfitting and enhancing the interpretability of the resulting models. However, it is essential to carefully choose the appropriate learning algorithm as the success of embedded methods heavily relies on this selection, and not all models or datasets are suitable for this approach.

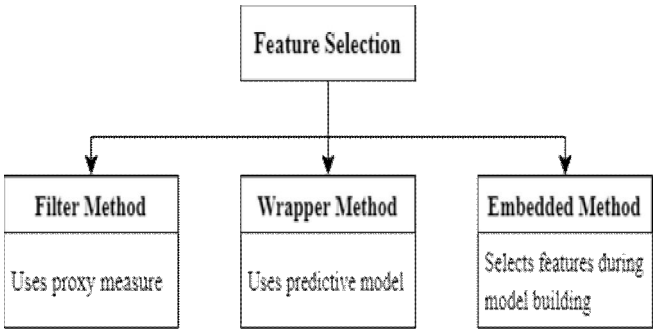


Fig -1: Types of Feature Selection methods and their main approach

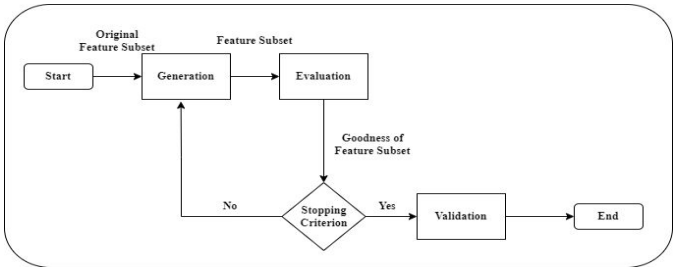


Fig -2: Feature Selection Procedure in General run-time.

B. Review on Feature Extraction Techniques

1) Feature extraction techniques play a crucial role in machine learning and computer vision by transforming raw data into meaningful and representative features. Dimensionality reduction methods, like Principal Component Analysis (PCA), map high-dimensional data to a lower-dimensional space, removing irrelevant or redundant information. PCA captures the most significant variations in the data, leading to improved computational efficiency and enhanced discriminative power. (Ref. Fig – 3).

2) Deep learning architectures, such as Convolutional Neural Networks (CNNs), are widely used for feature extraction in image-related tasks. CNNs extract hierarchical features by utilizing convolutional and pooling layers, capturing spatial and semantic information in input images. Transfer learning, which leverages pre-trained CNN models on large-scale datasets, enables effective feature extraction with limited labeled data, resulting in improved performance and faster convergence.

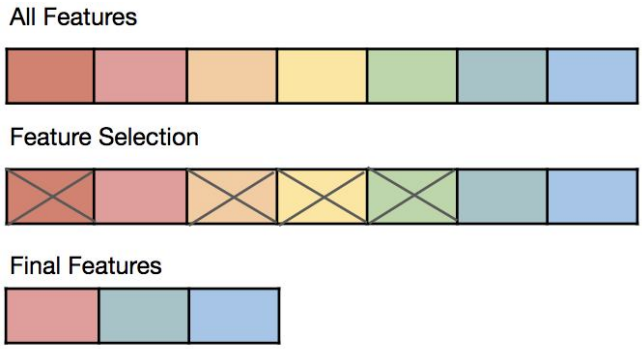


Fig -3: Feature Extraction Illustration

C. Review on Feature Selection and Extraction Techniques

Feature selection techniques and methods are fundamental in identifying relevant features and reducing dimensionality in machine learning tasks. Each category of feature selection approaches possesses distinct advantages and limitations, and the selection of the most suitable method depends on the specific requirements of the task, the characteristics of the dataset, and the available computational resources. Considering these factors carefully is crucial to ensure optimal feature selection, leading to improved model performance and interpretability.

Feature extraction techniques and methods are essential in machine learning and computer vision applications, enabling the transformation of raw data into meaningful representations. Dimensionality reduction techniques, such as PCA, help to reduce complexity and enhance discriminative power by retaining the most informative features. Deep learning-based methods, particularly CNNs and transfer learning, capture hierarchical and abstract features, enabling effective feature extraction from image data. Texture analysis techniques, including statistical and frequency-based approaches, provide valuable information about spatial patterns and enhance the accuracy of texture-related tasks. By employing appropriate feature extraction methods, researchers and practitioners can improve the performance and efficiency of various machine learning and image analysis tasks. Feature selection and evaluation techniques are crucial for improving model performance, interpretability, and efficiency.

D. Revew on Decision Tree in Image Classification

Decision trees have been widely utilized in image classification tasks, showcasing their effectiveness in certain contexts. With a hierarchical structure based on binary splits, decision trees offer interpretability and transparency in the decision-making process. They can handle both categorical and continuous features, making them versatile for image data.

Decision trees partition the feature space, capturing relationships between features and class labels. While decision trees can efficiently handle smaller datasets and provide insights into important features, they may struggle with capturing complex and nonlinear relationships in more intricate image datasets. Additionally, decision trees are prone to overfitting, particularly when the tree depth increases. Regularization techniques such as pruning can be employed to mitigate overfitting and improve generalization performance. Ultimately, the suitability of decision trees for image classification depends on the specific characteristics and complexity of the dataset at hand (Ref. Fig – 4).

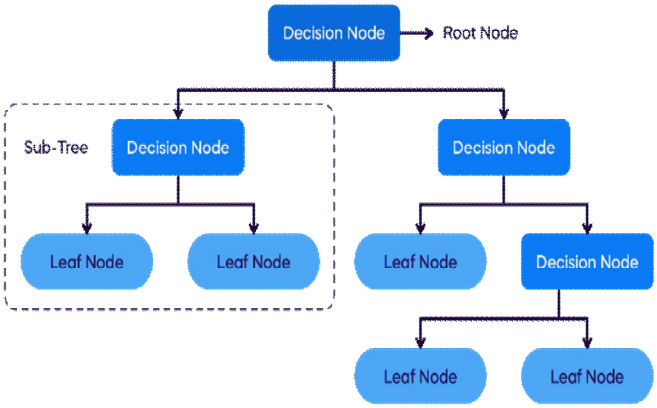


Fig -4: Decision Tree Learning Approach

III.METHODOLOGY

A. *Existing System*

1) Principal Component Analysis (PCA) is a widely used technique in machine learning and data analysis for dimensionality reduction. It can also be applied to feature selection in image classification tasks. PCA transforms the original features into uncorrelated variables known as principal components. These components capture the most significant variability in the data and can be ranked based on their variance. By selecting the top-ranked principal components, one can effectively reduce the dimensionality of the feature space while retaining important information, which is beneficial for image classification tasks.

2) Linear Discriminant Analysis (LDA) is another dimensionality reduction technique commonly employed for feature selection in image classification. Unlike PCA, LDA takes into account the class labels or target variable during the feature transformation process. By maximizing the separation between classes, LDA identifies a projection of the data that enhances class separability. This projection can highlight the most discriminative features for differentiating between classes, making LDA particularly useful in tasks like face recognition.

3) Mutual Information-based Feature Selection (MIFS) is a method that utilizes mutual information to assess the relevance of each feature to the target variable. Mutual information measures the statistical dependence between two random variables. In the context of feature selection, MIFS calculates the mutual information between each feature and the target variable, such as the class labels in image classification. Features with higher mutual information are considered more informative and are selected for the classification task. MIFS provides a principled approach to evaluate the information content of each feature in relation to the target, enabling the identification of the most relevant features for classification.

4) Recursive Feature Elimination (RFE) is an iterative feature selection method that aims to identify the most important features by eliminating the least significant ones. RFE starts with all features and employs a machine learning model to evaluate their importance. The model assigns weights or ranks to each feature based on its contribution to the prediction task. In each iteration, RFE eliminates the feature(s) with the lowest importance score and re-evaluates the remaining features. This iterative process continues until the desired number of features is reached. By iteratively removing the least important features, RFE focuses on retaining the most informative ones, potentially improving the classification performance and reducing overfitting.

B. *Proposed System*

The proposed system aims to make image classification better by using advanced techniques. It focuses on extracting more meaningful features from images, which can improve the accuracy of classification. The system also includes a smart way to select the best features and adjust its settings for better performance.

To extract better features, the system uses advanced methods that go beyond basic techniques. These methods help capture more relevant information from images, which can help in making more accurate classifications.

The system also includes a feature selection process called SelectFromModel. It automatically chooses the most important features from the extracted ones. This helps reduce the complexity of the classification process and makes it more efficient and another technique used is GridSearchCV. It fine-tunes the settings of the decision tree algorithm and ensemble classifier. By exploring

different combinations of settings, it finds the best configuration that improves classification performance. This helps the system make better decisions when classifying images.

In addition, the system uses ensemble classification, which means it combines the results of multiple classifiers. By doing this, it reduces biases and uncertainties that might be present in individual classifiers, leading to more reliable and accurate classification results.

The system also focuses on performance optimization. It adjusts the settings of the decision tree algorithm and ensemble classifier using SelectFromModel and GridSearchCV. This ensures that the system performs at its best, providing high accuracy and efficiency in classifying images (Ref. Fig – 5).

Overall, the proposed system follows a well-defined process. It involves collecting data, preparing it, extracting and selecting features using advanced techniques, training and adjusting the models, evaluating their performance, and finally using them for classification. This step-by-step approach ensures that the system effectively processes image data and produces improved classification outcomes (Ref. Fig – 6).

```
# Tuning HyperParameters
param_grid = {'max_depth': [10, 20, 30, 40], 'min_samples_split': [2, 4, 6, 8]}
grid_search = GridSearchCV(clf, param_grid=param_grid, cv=5)
grid_search.fit(X_train, y_train)
```

```
GridSearchCV(cv=5, estimator=DecisionTreeClassifier(random_state=42),
             param_grid={'max_depth': [10, 20, 30, 40],
                         'min_samples_split': [2, 4, 6, 8]})
```

```
print("Best parameters: ", grid_search.best_params_)
```

```
Best parameters:  {'max_depth': 20, 'min_samples_split': 2}
```

```
# Evaluate the feature importance
feature_importance = grid_search.best_estimator_.feature_importances_

# Select the most important features
model = SelectFromModel(grid_search.best_estimator_, prefit=True)
X_train_new = model.transform(X_train)
X_test_new = model.transform(X_test)
```

Fig -5: Proposed System – Code (Implementation)

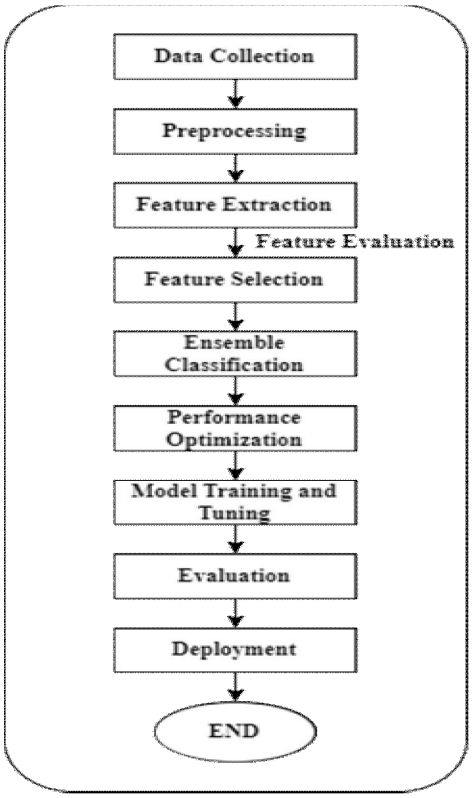


Fig -6: Proposed System Program flow

IV. RESULTS

Decision tree-based feature selection and feature extraction methods have great potential for advancing image classification. They can be applied to various fields like medical diagnosis, remote sensing, and surveillance systems. By further exploring and refining these methods, we can enhance the accuracy and efficiency of image classification, opening up new possibilities for practical applications (Ref. Fig – 7 & 8).

Example:

The below is an Example of the methodology proposed performed on sklearn-digits dataset:

- Accuracy on test set without selected features: 84.17%
- Accuracy on test set with selected features: 86.39%

Also Implementing on Various Datasets:

TABLE I: Comparison on datasets before and after performing feature evaluation and selection

Dataset Name	Accuracy	
	Before FE and FS	After FE and FS
1. sklearn-digits	84.17%	86.39%
2. mnist_784	65.79%	79.82%
3. fashion-mnist	71.10%	71.65%
4. Cifar-10	81.41%	87.55%
5. fruit data	83.33%	87.30%

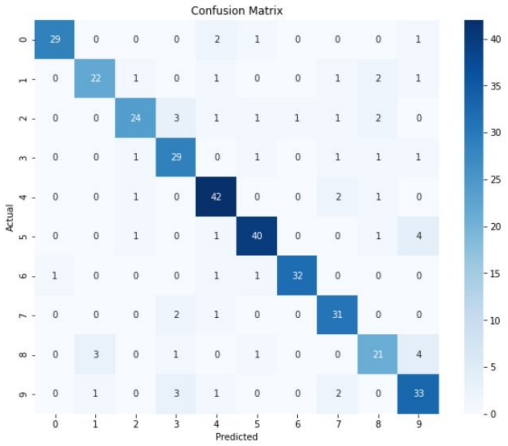


Fig -7: Confusion Matrix on sklearn digits dataset after feature selection and evaluation

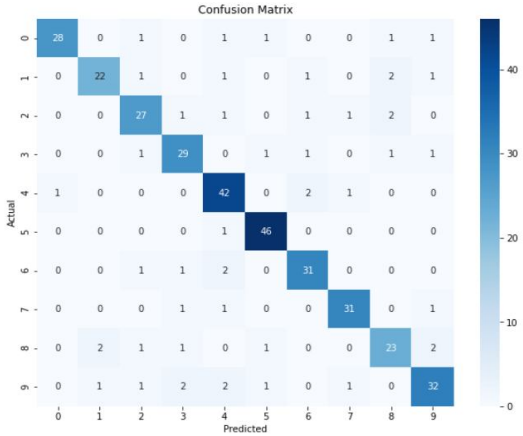


Fig -8: Confusion Matrix on sklearn digits dataset after feature selection and evaluation

V. CONCLUSION

In conclusion, decision tree-based feature selection and feature extraction methods have been shown to be effective for improving the performance of image classification. These methods use decision trees to select the most discriminative features or extract features from decision tree nodes, which can significantly reduce the dimensionality of the feature space and improve the accuracy of classification models.

Several studies have shown that decision tree-based feature selection and feature extraction methods outperform other methods in terms of accuracy and computational efficiency. Also, these methods can be easily integrated into existing classification models and require minimal pre-processing of the image data.

However, further research is needed to investigate the robustness of decision tree-based methods to variations in the image data, as well as their scalability to large datasets.

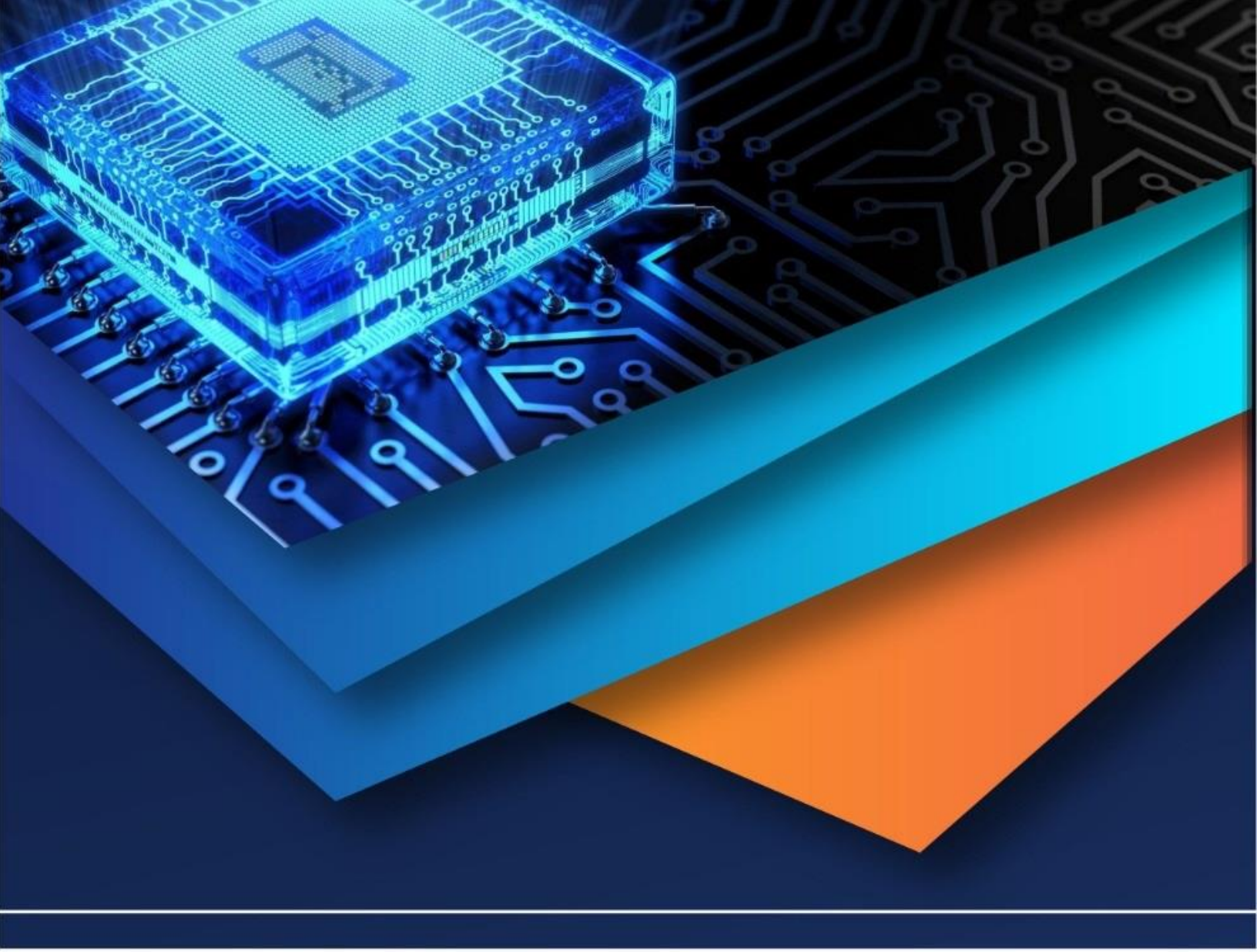
Overall, decision tree-based feature selection and feature extraction methods have great potential for advancing the field of image classification and can be applied to a wide range of applications, such as medical diagnosis, remote sensing, and surveillance systems.

VI. ACKNOWLEDGMENT

We would like to extend our sincere appreciation to Professor Yogesh Kakde[6], our project guide, and our Head of the Department (CSE - AI&ML) Dr. Thayyaba Khatoon, for their invaluable guidance, insightful feedback, and continuous support throughout this research. Lastly, we would like to thank our families, friends, and loved ones for their unwavering support and encouragement during the completion of this research.

REFERENCES

[1] I. Guyon, “An introduction to variable and feature selection,” Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.
[2] P. Langley, “Selection of relevant features in machine learning,” in Proceedings of the AAAI Fall Symposium on Relevance. Washington, D.C., USA: AAAI Press, 1994, pp. 127–131.
[3] L. Yu and H. Liu, “Feature selection for high-dimensional data: A fast correlation-based filter solution,” in Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington, D.C., USA, 21-24 August 2003, pp. 856–863.
[4] P. W. Frey and D. J. Slate, “Letter recognition using holland-style adaptive classifiers,” Machine Learning, vol. 6, no. 2, pp. 161–182, 1991.
[5] M. Dash and H. Liu, “Feature selection for classification,” Intelligent Data Analysis, vol. 1, pp. 131–156, 1997.
[6] M. A. Hall and L. A. Smith, “Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper,” in Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference, Orlando, Florida, 1-5 May 1999, pp. 235– 239.
[7] Han Liu, Mihaela Cocea and Weili Ding, “Decision tree learning based feature evaluation and selection for image classification,” in International Conference on Machine Learning and Cybernetics (ICMLC), Ningbo, China, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY