Karadeniz Fen Bilimleri Dergisi, 15(1), 560-576, 2025. DOI: 10.31466/kfbd.1620640

KFBD

Karadeniz Fen Bilimleri Dergisi
The Black Sea Journal of Sciences
ISSN (Online): 2564-7377 https://dergipark.org.tr/tr/pub/kfbd

Araştırma Makalesi / Research Article

Vision Transformer-Based Approach: A Novel Method for Object Recognition

Ali Khudhair Abbas ALI[1], Yıldız AYDIN[2*]

Abstract

This paper proposes a hybrid method to improve object recognition applications on inefficient and imbalanced datasets. The proposed method aims to enhance object recognition performance using the Vision Transformer (ViT) deep learning model and various classical machine learning classifiers (LightGBM, AdaBoost, ExtraTrees, and Logistic Regression). The Caltech-101 dataset used in the study is a low-resolution and noisy image dataset with class imbalance problems. Our method achieves better results by combining the feature extraction capabilities of the Vision Transformer model and the robust classification performance of classical machine learning classifiers. Experiments conducted on the Caltech-101 dataset demonstrate that the proposed method achieves a precision of 92.3%, a recall of 89.7%, and an accuracy of 95.5**%**, highlighting its effectiveness in addressing the challenges of object recognition in imbalanced datasets.

Keywords: Object recognition, Vision Transformer, Logistic Regression, Caltech 101, Image Processing, Artificial Intelligence.

Görsel Dönüştürücü Tabanlı Yaklaşım: Nesne Tanıma için Yeni Bir Yöntem

Öz

Bu makale, verimsiz ve dengesiz veri kümeleri üzerinde nesne tanıma uygulamalarını iyileştirmek için hibrit bir yöntem önermektedir. Önerilen yöntem, Vision Transformer (ViT) derin öğrenme modelini ve çeşitli klasik makine öğrenimi sınıflandırıcılarını (LightGBM, AdaBoost, ExtraTrees ve Logistic Regression) kullanarak nesne tanıma performansını artırmayı amaçlamaktadır. Çalışmada kullanılan Caltech-101 veri kümesi, sınıf dengesizliği sorunları olan düşük çözünürlüklü ve gürültülü bir görüntü veri kümesidir. Yöntemimiz, Vision Transformer modelinin özellik çıkarma yetenekleri ile klasik makine öğrenimi sınıflandırıcılarının sağlam sınıflandırma performansını birleştirerek daha iyi sonuçlar elde etmektedir. Caltech-101 veri kümesi üzerinde yapılan deneyler, önerilen yöntemin %92,3'lük bir hassasiyet ve %89,7'lik bir geri çağırma elde ettiğini ve diğer son teknoloji yöntemlerden önemli ölçüde daha iyi performans gösterdiğini ortaya koymaktadır. Deneysel sonuçlar, önerilen yöntemin diğer mevcut yöntemlerden daha iyi performans gösterdiğini ve nesne tanıma görevlerinde önemli iyileştirmeler sağladığını göstermektedir.

Anahtar Kelimeler: Nesne tanıma, Vision Transformer, Lojistik Regresyon, Caltech 101, Görüntü İşleme, Yapay Zeka.

[1]Institute of Science and Technology, Erzincan Binali Yıldırım University, Erzincan, Türkiye, alialsodane567@gmail.com
[2]Department of Computer Engineering, Erzincan Binali Yıldırım University, Erzincan, Türkiye, yciltas@erzincan.edu.tr

[*]Sorumlu Yazar/Corresponding Author

Geliş/Received: 15.01.2025 **Kabul/Accepted:** 05.03.2025 **Yayın/Published:** 15.03.2025

Ali, A.K.A., Aydın, Y. (2025). Vision Transformer-Based Approach: A Novel Method for Object Recognition. *Karadeniz Fen Bilimleri Dergisi*, 15(1), 560-576.

1. Introduction

Object recognition is a frequently researched topic in artificial intelligence and image processing. Applications that identify objects found in images are called object recognition applications. Object recognition applications, which are an important subcomponent in smart robot production, are also used in security systems, automatic driving technologies, and health fields (Dosovitskiy et al., 2021; Krizhevsky, Sutskever, & Hinton, 2012). With the rapid development of technology, the use of smart devices in daily life is increasing day by day. Therefore, it is very important to develop an effective and efficient object recognition application against variable factors such as rotation, scale, and illumination changes. Objects in images used in real-world applications may not always be clear and problems such as background clutter may be encountered. An effective and efficient solution needs to be developed to overcome such problems in object recognition applications. While artificial neural networks and deep learning-based approaches achieve high accuracy, they also present challenges such as class imbalance and high computational cost (Fei-Fei, Fergus, & Perona, 2006). Recently, Vision Transformer (ViT) models have demonstrated significant success in object recognition tasks, emerging as an alternative to traditional CNN-based methods (Touvron et al., 2021). However, the high computational requirements of ViT models and their performance decline on imbalanced datasets have necessitated the development of novel hybrid approaches (Naseer, Alzahrani, et al., 2024). This study aims to enhance object recognition performance by integrating ViT-based feature extraction with classical machine learning classifiers.

Object recognition applications basically consist of two steps: feature extraction and classification. The methods to be used in each of these steps directly affect the success of the application. Object recognition applications can be developed with classical machine learning classifiers and deep learning methods. Feature extraction is a crucial step in object recognition and image processing applications. The feature extraction methods used in this study have distinct structural and algorithmic properties. SIFT (Scale-Invariant Feature Transform) is widely used for object recognition due to its robustness against scale and rotation changes. SURF (Speeded-Up Robust Features) is similar to SIFT but offers improved computational efficiency. ORB (Oriented FAST and Rotated BRIEF) is a fast and rotation-invariant feature extraction technique. KAZE detects edges and textures using differential operators, while MSER (Maximally Stable Extremal Regions) identifies regions that remain stable under varying brightness conditions. BRISK (Binary Robust Invariant Scalable Keypoints) is a fast and scalable feature extraction method. These techniques help identify key points in an image, enhancing the performance of classification algorithms. While local features or global features such as SIFT, SURF, KAZE, and ORB are used in classical machine learning methods, deep learning methods extract the features themselves.

The main problems that may be encountered in object recognition applications are:

1.The object in the image blends with the background.

2.Low-resolution or blurry images.

3.Variability of lighting conditions.

4.Rotating, scaling, or partially obstructing the object.

Various hybrid approaches are proposed in the literature to solve the problems encountered in object recognition applications (Naseer, Almujally, Alotaibi, Alazeb, & Park, 2024; Naseer, Mudawi, et al., 2024; Sikder, Islam, & Jahan, 2024). The successful results obtained with the use of hybrid methods in recent years have led to an increase in the number of studies on hybrid methods. Hybrid systems are obtained by combining deep learning methods, classical features, or classical machine classifiers with different variations. With these hybrid systems, it has been possible to obtain more successful results in systems with low success rates, especially in applications developed with classical machine learning or deep learning methods. Examples of these applications are those performed with an unbalanced dataset.

In unbalanced datasets, while the number of samples in one class is quite low, the number of samples in other classes can be quite high. It is very important to solve this problem, which is frequently encountered in daily life. In this study, the unbalanced Caltech 101 dataset was used. It has been observed in studies in the literature that in applications carried out with classical methods using unbalanced datasets, there is a tendency towards the class with a large number of samples. In this study, a hybrid method is proposed to develop a more effective and efficient object recognition application by eliminating the problems of unbalanced datasets. In the proposed hybrid method, the Vision Transformer (ViT) method was used in the feature extraction step, and LightGBM, AdaBoost, ExtraTrees, and Logistic Regression classifiers were used in the classification step. The main motivation for adopting this approach is based on two main reasons:

- To overcome the limitations of imbalanced datasets, deep learning-based features can be integrated with robust classification models, enabling better generalization, especially for underrepresented classes.
- To increase recognition accuracy on noisy and low-resolution datasets without requiring large-scale computational resources, making the method more efficient and applicable in real-world applications.

In order to combine the strengths of traditional machine learning and deep learning approaches, the proposed method integrates the Vision Transformer (ViT) model with classical machine learning classifiers. ViT extracts deep features with a self-attention mechanism, while these features are classified by powerful machine learning algorithms such as LightGBM, AdaBoost, ExtraTrees, and Logistic Regression. This hybrid approach overcomes the limitations of deep learning, such as high

computational requirements and sensitivity to class imbalance, while improving the classification accuracy. Instead of using only deep learning-based classification, the proposed method improves the generalization of underrepresented classes, especially by processing the extracted features more efficiently using machine learning classifiers. The ViT method was preferred in the feature extraction step because this deep learning-based method can achieve more successful results with fewer parameters (Venugopal, Joseph, Vipin Das, & Kumar Nath, 2022).

As a result, a more effective and efficient object recognition application has been achieved with the proposed hybrid method.

Main Contributions of the Proposed Method to the Literature

- Elimination of the imbalance problem on the Caltech101 dataset, which is an unbalanced dataset.
- A more successful hybrid approach is proposed by using features extracted with deep learning methods and classical machine learning classifiers.
- The method has been empirically validated through extensive experiments, demonstrating superior performance over existing state-of-the-art methods in object recognition tasks.

The main innovation of the proposed method is to eliminate the disadvantages of imbalanced datasets where certain object categories are not sufficiently represented. Most of the existing studies eliminate these disadvantages by applying data augmentation techniques or ensemble learning strategies to eliminate class imbalance. However, the proposed method offers a more effective solution by integrating the deep feature extraction method with classical machine classifiers. In the proposed approach, the Caltech-101 dataset, which is an imbalanced dataset widely used in the literature, is used and experimental results show that the proposed approach outperforms the state-of-the-art methods in terms of precision (92.3%), recall (89.7%) and accuracy (95.5%). In the second part of the article, the relevant literature is detailed, and in the third part, the method is detailed. In Section 4, the results obtained from the experiments are given. Finally, the results and recommendations are explained in 4 chapters.

2. Related Work

Image recognition is an important research topic in computer vision and has gained a new dimension in recent years with the widespread use of Convolutional Neural Network (CNN) models (KARADAĞ & ÖZDEMİR, 2022; Keerthana, Venugopal, Nath, & Mishra, 2023; Telceken & Kutlu, 2022). CNNs have shown superior performance in image recognition problems compared to traditional machine learning methods (R. Zhang, Wang, Cheng, & Song, 2023). For example, the AlexNet study by Krizhevsky et al. (Krizhevsky et al., 2012) proved the power of deep learning with

groundbreaking results on the ImageNet dataset.

The Vision Transformer (ViT) model is a new approach that uses the mechanism of self-attention to understand correlations in visual data. When Dosovitskiy et al. (Dosovitskiy et al., 2021) introduced ViT, they noted that this model requires large-scale data sets and high computational power to achieve high performance. ViT has been particularly successful in image recognition tasks, but its high computational costs limit its application areas.

In order to develop more computable and efficient models, hybrid approaches that combine ViT with other machine learning methods have been proposed. For example, the DeiT (Data-efficient Image Transformers) model developed by Touvron et al. (Touvron et al., 2021) uses distillation to speed up the training process of ViT and achieve high performance with less data.

Bosh et al. (Bosch, Zisserman, & Muñoz, 2007) suggested a technique for identifying several object types using ExtraTrees classifiers. Using the Caltech101 and Caltech 256 datasets, they demonstrated that their recommended approach, which included ROI and other adjustments, improved performance by 5%.

In the study conducted by Liu et al. (Liu, Guo, Chamnongthai, & Prasetyo, 2017) a more successful system was proposed by combining the color histogram with Local Binary Patterns (LBP). Since the LBP feature does not effectively utilize the color information of images, they combined it with the Color Information Feature (CIF) to leverage color features. Experimental results showed that this hybrid feature was more effective. Bansal et al. (Bansal, Kumar, & Kumar, 2021) used SIFT, SURF, and ORB feature descriptors in their object recognition application and comparatively analyzed the performance of these three methods. They examined the performance of SIFT, SURF, and ORB features separately, as well as hybrid features obtained by combining these descriptors in different ways. They reported that the method combining SIFT, SURF, and ORB features achieved the best performance.

Naseer et al. (Naseer, Almujally, et al., 2024) utilized a segmentation approach in combination with Artificial Neural Networks (ANNs). First, they segmented the input images using the RGB color space. These segmented images were then classified using ANNs, producing accuracy rates of 89%, 83%, and 90% on the MSRC, MS COCO, and Caltech 101 datasets, respectively.

Hussein et al. (Hussain et al., 2024) proposed three steps for object recognition. The first stage involves data augmentation to address the imbalance issue in the Caltech 101 dataset. The second stage combines features from Inception V3 deep learning, Pyramid HOG (PHOG), and Central Symmetric LBP (CS-LBP). The final step uses the Joint Entropy and KNN (JEKNN) technique to determine the optimal features. Their method achieved an accuracy rate of 90.4% on the Caltech 101 dataset.

Zhang et al (L. Zhang & Pu, 2024) developed a hybrid model that combines Hu moments and Speeded up Robust Feature (SURF) methods for object recognition tasks. The two different methods are effectively combined using various weighting factors. Their proposed method is shown to be robust to changes in scale, viewpoint, illumination and noise, and more effective and robust than other representative methods.

Most of the above studies do not include traditional models in their evaluation and propose multi-stage methods to address the imbalance problem. While various techniques such as ensemble methods can address this issue, more robust features can also provide a solution. The proposed ViT and logistic regression-based method overcomes this problem. Our proposed method outperforms existing state-of-the-art methods in scene recognition and object classification, making a significant contribution to the literature.

3. The Suggested Object Recognition Approach

In recent years, the use of fusion methods has become increasingly popular in artificial intelligence applications. Some approaches involve the fusion of features, while others combine deep learning and machine learning methods. Although there is a growing body of work in the literature that integrates SIFT features with classical machine classifiers and CNN deep learning methods, there remains a notable gap in fusion-based approaches utilizing alternative techniques.

This research presents a hybrid object identification approach that combines classical machine learning classifiers for the classification stage with the feature extraction strengths of ViT. This approach aims to address and overcome common issues in object identification tasks, particularly those involving imbalanced datasets. ViT was selected for this study because it delivers superior results with fewer parameters compared to other deep learning models. (Tan & Le, 2019). In addition, ViT is a robust feature extractor under various conditions such as rotation, scale, and illumination changes. In datasets like Caltech101, which is imbalanced, it is crucial to select a robust feature extractor to prevent instability issues. To address the imbalance inherent in datasets like Caltech101, the proposed method combines features extracted by ViT with classical machine learning classifiers, including LightGBM, AdaBoost, ExtraTrees, and Logistic Regression. By integrating these classifiers, the method not only leverages the powerful feature extraction capabilities of deep learning but also benefits from the robust classification performance of classical techniques.

The ViT and Logistic Regression methods used in the proposed approach are explained in detail below. The applications conducted as part of this study are illustrated in Figure 1.

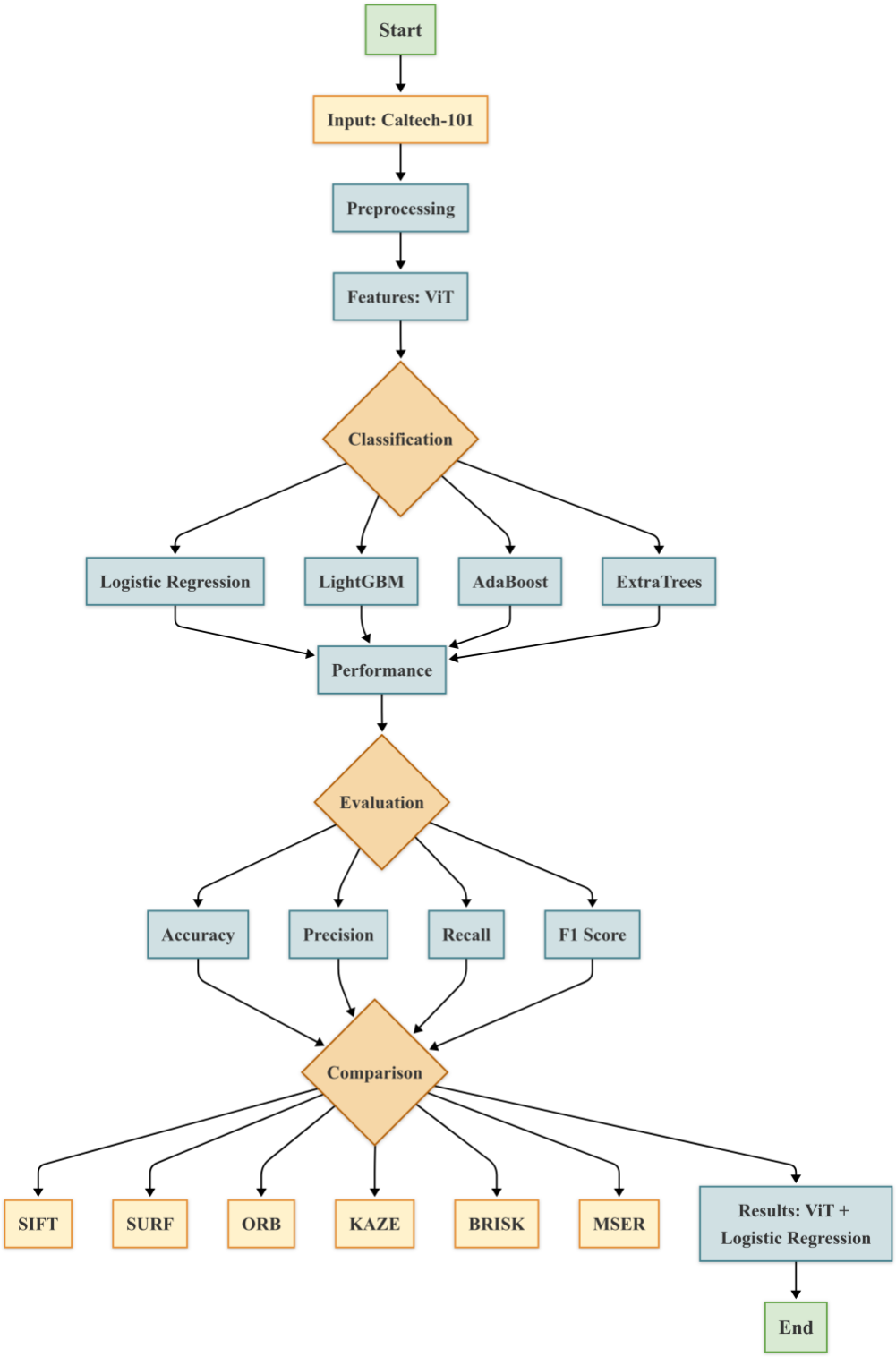

Figure 1. Flow diagram showing the general structure of the study.

The proposed method extracts deep features using the Vision Transformer (ViT) model and evaluates these features with machine learning classifiers such as LightGBM, AdaBoost, ExtraTrees, and Logistic Regression. To balance the high computational requirements of deep learning models like ViT, the extracted features were tested using Stratified K-Fold cross-validation. Each classifier's

accuracy, F1-score, and classification reports were calculated to conduct a comprehensive performance analysis. All classifiers were used with their default parameter settings.

The proposed method was implemented and tested using Google Colab, which provides cloud-based GPU resources. The Vision Transformer (ViT) model, while computationally intensive, was efficiently utilized within Colab's hardware constraints. The main computational burden lies in the feature extraction stage, as ViT processes image patches with a self-attention mechanism. However, once features are extracted, classification is performed using lightweight machine learning models such as Logistic Regression, LightGBM, AdaBoost, and ExtraTrees, significantly reducing inference time and resource usage. This approach enables the method to be adaptable for deployment on devices with limited computational power by leveraging pre-extracted features, making it suitable for real-world applications with constrained hardware environments.

3.1. Vision Transformer (ViT)

The Vision Transformer (ViT) model, which can perform particularly well on large datasets, can effectively model the global context in the image (Dosovitskiy et al., 2021). In the ViT model, the input images are processed as an index by segmenting them into small patches of fixed size. Equation 1 gives the total number of P −dimensional patches obtained from the input image.

$$N = \frac{H \times W}{P^2} \tag{1}$$

It's here, H is the height and W is the width of the image. After each patch is regularized as a vector of dimension x_i, it is embedded into a lower dimensional space through a linear layer (Equation (2)).

$$z_0^i = E \cdot x_i + e_{pos}^i, for\ i = 1, \dots, N \tag{2}$$

In Equation 2, E is a learnable embedding matrix, z_0^i is the embedded vector of the i-th patch and e_{pos}^i is the positional encoding. Each embedded vector is processed by the Transformer Encoder. The encoder layer consists of multi-head attention and feed-forward network items (Equation 3).

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

Where Q, K and Z are the query, key, and value matrices respectively. d_k is the (head) dimension and each encoder layer produces the process given in Equation 4. MLP is a feed-forward neural network and z_l is the input vector from the l-th layer.

$$z_{l+1} = MLP\big(Attention(z_l)\big) + z_l \tag{4}$$

The final output from all Transformer Encoder layers is transmitted to the classification layer, which consists of a fully connected network, and the class labels are obtained by the mathematical expression given in equation 5.

$$\hat{y} = softmax(W_{fc} \cdot z_l) \quad (5)$$

In Equation 5, W_{fc} is the weight matrix used for classification and z_l is the output vector from the last Encoder layer. The cross entropy loss function is used to train the model and this function is optimized to minimize the difference between the predicted class distribution and the actual class distribution.

3.2. LogisticRegression

Logistic regression, a method of multivariate regression analysis, is widely applied in various fields such as medicine, marketing, finance, and more (Jin et al., 2022). The main purpose of logistic regression is to estimate the probability that the dependent variable belongs to a particular class. This model uses a logistic function (also known as the sigmoid function) to estimate these probabilities. The basic formula for logistic regression is given by Equation (6).

$$P = \frac{1}{1+e^{-z}} \quad (6)$$

Where P is the probability of the event occurring and z is defined as Equation (7).

$$Z = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (7)$$

β_0 is the intercept (bias) of the model, and β_{n} are the coefficients of the independent variables.

The logistic function calculates the probability of an event occurring, ensuring that the output is between 0 and 1. The sigmoid function in logistic regression transforms a linear equation into a probability between 0 and 1. This transformation enhances the model's sensitivity to changes in the independent variables and allows it to capture non-linear relationships.

4. Experimental Results

The Caltech-101 dataset (Fei-Fei, Fergus, & Perona, 2004; Mutch & Lowe, 2006), commonly used in object classification problems, presents a challenging multi-class structure. Comprising 9,146 images across 102 categories, it includes a background scene category, which has been excluded in some studies due to its diversity. The number of images per category ranges from at least 31 to a maximum of 800, with images being low-resolution and noisy. The objective of using the Caltech-101 dataset in this study is to leverage robust features to address the challenges posed by this unstable and difficult dataset.

In the proposed method, feature extraction is performed using the deep learning technique ViT, while classification is carried out using advanced methods such as LightGBM, AdaBoost, ExtraTrees, and Logistic Regression. Various performance metrics are employed to compare feature extraction

methods like SIFT, SURF, ORB, KAZE, MSER, and BRISK. Given the multi-class nature of the dataset, the macro-averaging method is used to evaluate performance by averaging the prediction results across all classes. Four parameters—accuracy, precision, recall, and F1 score—are used to analyze the experimental results. Since the Caltech-101 dataset is not divided into training and test sets, cross-validation, a standard evaluation methodology, is employed. The experiments also include a comparative analysis of various advanced classifiers (LightGBM, AdaBoost, ExtraTrees, and Logistic Regression). Table 1 presents a comparison of feature extraction methods based on recognition accuracy, while Figure 3 graphically illustrates the performance of the proposed system.

Table 1. Precision values for various features and classifiers.

Features	LightGBM	AdaBoost	ExtraTrees	Logistic Regression
SIFT	0.2533	**0.0438**	0.2154	0.3138
SURF	0.2785	0.0366	0.2975	0.2691
ORB	0.1958	0.0150	0.1531	0.2355
KAZE	0.3092	0.0327	0.3763	0.4187
BRISK	0.1625	0.0190	0.1274	0.1651
MSER	0.2115	0.0228	0.2439	0.1864
Proposed Method	**0.4399**	0.0331	**0.9506**	**0.9566**

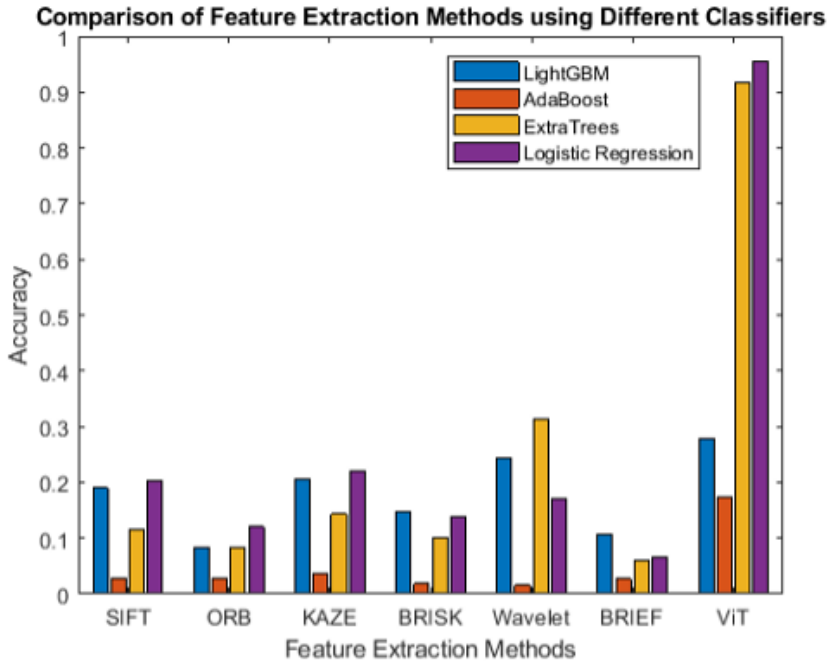

Figure 2. Accuracy values of methods

Table 2 and Table 3 present the comparison based on precision and recall. These comparative results are illustrated graphically in Figure 3, which provides a clear depiction of the performance of the proposed system.

It has been observed from the experimental results that significant improvements were achieved when ExtraTrees and Logistic Regression were employed in the proposed method. The robust features extracted by the Vision Transformer were effectively utilized by these classifiers, resulting in markedly enhanced precision, recall, and F1-score values. In contrast, suboptimal results were recorded for LightGBM and AdaBoost, which are believed to have been adversely affected by their sensitivity to class imbalance and less efficient handling of the deep features. Hence, the superiority of the proposed method is attributed to the optimal integration of deep feature extraction and classical classification techniques. As shown in all the tables, the results obtained from the proposed approach demonstrate improvements across all performance metrics. Figure 2 and Figure 3 provide a detailed comparison of the various classifiers on all performance parameters, indicating that the proposed fusion approach is more advantageous than the classical methods.

Table 2. Recall values for various features and classifiers

Features	**LightGBM**	**AdaBoost**	**ExtraTrees**	**Logistic Regression**
SIFT	0.1669	0.0347	0.0897	0.2588
SURF	0.1418	0.0248	0.0970	0.2086
ORB	0.1022	0.0208	0.0547	0.1648
KAZE	0.2498	0.0302	0.1580	0.3554
BRISK	0.0842	0.0217	0.0557	0.1276
MSER	0.1274	0.0327	0.1032	0.1490
Proposed-Method	**0.2578**	**0.0347**	**0.8946**	**0.9393**

Table 3. F1-scores for various features and classifiers

Features	**LightGBM**	**AdaBoost**	**ExtraTrees**	**Logistic Regression**
SIFT	0.1792	**0.0272**	0.0912	0.2749
SURF	0.1592	0.0175	0.1020	0.2255
ORB	0.1119	0.0139	0.0501	0.1824
KAZE	0.2630	0.0203	0.1741	0.3747
BRISK	0.0867	0.0134	0.0515	0.1345
MSER	0.1318	0.0219	0.0995	0.1575
Proposed-Method	**0.2988**	0.0266	**0.9181**	**0.9474**

All experiments were conducted on a machine running Microsoft Windows 10 (genuine) with an Intel Core i5 processor and 8 GB of RAM.

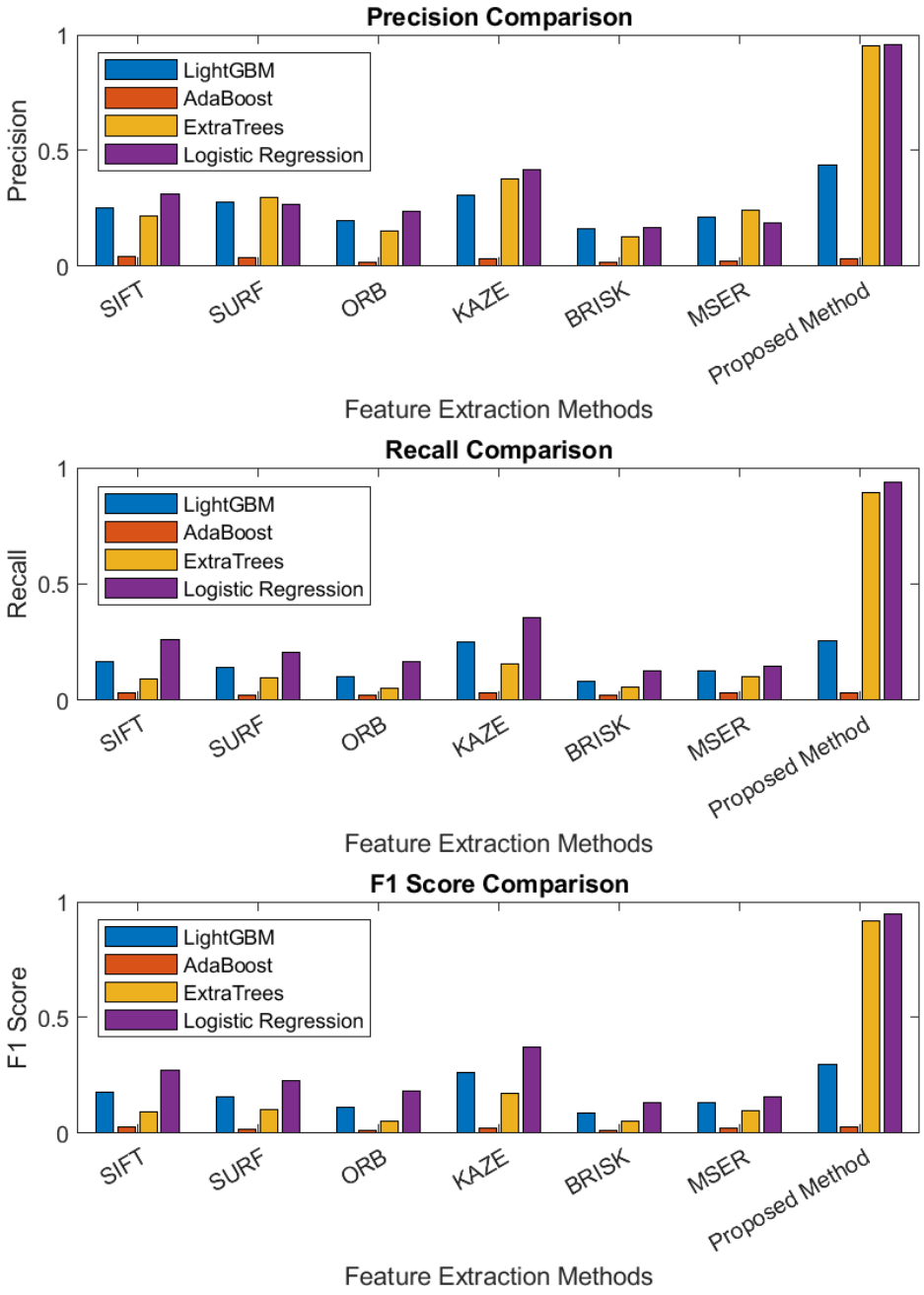

Figure 3. Precision, recall, F1 values of methods

Table 1 shows the accuracy values obtained by different feature extraction methods with various classifiers, while Tables 2 and 3 provide detailed comparisons in terms of precision, recall and F1 scores. It is seen that ViT-based features in particular provide more consistent and higher success compared to other traditional methods. In Figures 2 and 3, the improvements provided by the proposed method in terms of different metrics are graphically presented and it is seen that our method is significantly superior to traditional methods. This success can be explained by the strong feature

extraction ability of ViT and the optimized classification process with classical machine learning classifiers. The comparisons clearly reveal that the proposed method provides more reliable and successful results on low-quality and imbalanced datasets. The superior performance of the proposed method compared to traditional approaches is attributed to its robust feature extraction and efficient classification strategy. Unlike handcrafted feature extraction techniques such as SIFT, SURF, and ORB, which are sensitive to variations in scale, rotation, and noise, the Vision Transformer (ViT) is capable of capturing global contextual information using a self-attention mechanism. This results in more discriminative features, which are particularly beneficial for imbalanced and noisy datasets like Caltech-101. Instead of relying solely on deep learning for classification, ViT-extracted features are combined with classical machine learning classifiers such as Logistic Regression, LightGBM, AdaBoost, and ExtraTrees, ensuring computational efficiency without compromising accuracy. Traditional classifiers often face challenges with class imbalance, leading to biased predictions; however, this issue is mitigated by leveraging ViT's powerful feature representations, which enhance generalization across all classes. Furthermore, as the experiments were conducted on Google Colab, the feasibility of the method is demonstrated without the need for high-end hardware. The analysis of Figures 2 and 3 confirms that, while traditional methods exhibit inconsistent performance across different classifiers, the proposed approach maintains superior accuracy, precision, and recall, making it a highly effective solution for object recognition tasks.

Recently, various researchers have explored different ensemble approaches for image classification due to their improved accuracy results. Table 4 presents a comparative analysis of the proposed system alongside some recent experiments conducted on the Caltech-101 dataset.

Table 4. Comparative analysis of the proposed system with some recent experiments

Method	**Accuracy**
Jalal et al. (Jalal, Ahmed, Rafique, & Kim, 2021)	0.8926
Naseer et al.(Naseer, Almujally, et al., 2024)	0.9030
Rafique et al. (Rafique, Gochoo, Jalal, & Kim, 2023)	0. 8860
Hussain et al. (Hussain et al., 2024)	0.9040
Gupta et al. (Gupta, Kumar, & Garg, 2019)	0.8560
Bansal et al. (Bansal, Kumar, Kumar, & Kumar, 2021)	0.8640
Bansal et al. (Bansal, Kumar, & Kumar, 2021)	0.8330
Proposed Method	**0.9550**

This study makes a significant contribution to the literature through an in-depth analysis and comparison using the Caltech-101 dataset. The proposed method and the results obtained highlight the effectiveness of robust feature extraction and advanced classification techniques in addressing

object classification problems. In this context, this study provides a strong foundation for future research and offers valuable insights for subsequent studies.

5. Conclusions

This study aims to address the imbalance problem by employing powerful feature extraction and classification methods on imbalanced datasets. To this end, features extracted using the ViT model are classified using Logistic Regression, LightGBM, ExtraTrees, and AdaBoost methods, which are commonly used classifiers for imbalanced datasets. The experimental results demonstrate that features extracted with the ViT model when combined with the Logistic Regression classifier, achieve high accuracy rates on imbalanced datasets.

Analysis of accuracy, precision, recall, and F1-score metrics reveals that the Logistic Regression and ExtraTrees classifiers are less affected by the imbalance problem and provide effective solutions. Specifically, Logistic Regression achieved high success rates, whereas LightGBM and AdaBoost classifiers were adversely affected by the imbalance problem, resulting in lower success rates.

Experiments on the Caltech-101 dataset show that the proposed hybrid method surpasses both classical and state-of-the-art methods. Comparisons with feature extraction methods such as SIFT, SURF, KAZE, ORB, MSER, and BRISK indicate that the ViT and Logistic Regression-based approach delivers superior accuracy, precision, recall, and F1 score. The proposed method effectively mitigates the imbalance problem, offering a more successful hybrid approach by combining features extracted through deep learning with classical machine learning classifiers, and outperforming existing advanced methods. Although the proposed approach demonstrates superior performance in object recognition, certain limitations must be acknowledged. One of the main constraints is the computational complexity introduced by the Vision Transformer's feature extraction process, which may limit its applicability in resource-constrained environments. Additionally, the sensitivity of some classifiers, particularly LightGBM and AdaBoost, to class imbalance has been observed, potentially leading to biased predictions in highly imbalanced datasets. Future research will focus on addressing these limitations by exploring more efficient feature extraction techniques, adaptive classification strategies, and automated hyperparameter tuning methods to further improve the robustness and scalability of the proposed approach.

The proposed method can increase the effectiveness of systems by improving the detection of rare or anomalous events in health and safety applications thanks to its ability to eliminate the disadvantages of imbalanced datasets. In addition, the proposed method can be applied in real-time object recognition in autonomous systems, especially for driverless vehicles and robotic navigation,

thanks to its robustness to illumination, scale and noise changes. In conclusion, this study provides an effective and efficient solution to the imbalanced dataset problem and outperforms recent studies in the literature. Future work will focus on developing applications with different datasets and exploring more complex hybrid methods.

Authors' Contributions

All authors contributed equally to the study.

Statement of Conflicts of Interest

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The author declares that this study complies with Research and Publication Ethics.

References

Amerini, I., Ballan, L., Caldelli, R., Del Bimbo, A., & Serra, G. (2011). A SIFT-based forensic method for copy-move attack detection and transformation recovery. *IEEE Transactions on Information Forensics and Security*, *6*(3 PART 2), 1099–1110. https://doi.org/10.1109/TIFS.2011.2129512

Bansal, M., Kumar, M., & Kumar, M. (2021). 2D object recognition: a comparative analysis of SIFT, SURF and ORB feature descriptors. *Multimedia Tools and Applications*, *80*(12), 18839–18857. https://doi.org/10.1007/s11042-021-10646-0

Bansal, M., Kumar, M., Kumar, M., & Kumar, K. (2021). An efficient technique for object recognition using Shi-Tomasi corner detection algorithm. *Soft Computing*, *25*(6), 4423–4432. https://doi.org/10.1007/s00500-020-05453-y

Bosch, A., Zisserman, A., & Muñoz, X. (2007). Image classification using random forests and ferns. *Proceedings of the IEEE International Conference on Computer Vision*. https://doi.org/10.1109/ICCV.2007.4409066

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2021). an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale. *ICLR 2021 - 9th International Conference on Learning Representations*.

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Pattern Recognition Workshop*, 178. https://doi.org/10.1016/j.cviu.2005.09.012

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4), 594–611. https://doi.org/10.1109/TPAMI.2006.79

Gupta, S., Kumar, M., & Garg, A. (2019). Improved object recognition results using SIFT and ORB feature detector. *Multimedia Tools and Applications*, *78*(23), 34157–34171. https://doi.org/10.1007/s11042-019-08232-6

Hussain, N., Khan, M. A., Sharif, M., Khan, S. A., Albesher, A. A., Saba, T., & Armaghan, A. (2024). A deep neural network and classical features based scheme for objects recognition: an application for machine inspection. *Multimedia Tools and Applications*, *83*(5), 14935–14957. https://doi.org/10.1007/s11042-

020-08852-3

Jalal, A., Ahmed, A., Rafique, A. A., & Kim, K. (2021). Scene Semantic Recognition Based on Modified Fuzzy C-Mean and Maximum Entropy Using Object-to-Object Relations. *IEEE Access*, *9*, 27758–27772. https://doi.org/10.1109/ACCESS.2021.3058986

Jin, J., Chen, G., Meng, X., Zhang, Y., Shi, W., Li, Y., … Jiang, W. (2022). Prediction of river damming susceptibility by landslides based on a logistic regression model and InSAR techniques: A case study of the Bailong River Basin, China. *Engineering Geology*, *299*(February). https://doi.org/10.1016/j.enggeo.2022.106562

KARADAĞ, C., & ÖZDEMİR, D. (2022). *BÖBREK TÜMÖRÜ TESPİTİ İÇİN DERİN ÖĞRENME YÖNTEMLERİNİN KARŞILAŞTIRMALI ANALİZİ*. *6*(6), 10–23.

Keerthana, D., Venugopal, V., Nath, M. K., & Mishra, M. (2023). Hybrid convolutional neural networks with SVM classifier for classification of skin cancer. *Biomedical Engineering Advances*, *5*(December 2022), 100069. https://doi.org/10.1016/j.bea.2022.100069

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, *25↓*, V2-257-V2-259. https://doi.org/10.1016/B978-0-12-374105-9.00493-7

Liu, P., Guo, J. M., Chamnongthai, K., & Prasetyo, H. (2017). Fusion of color histogram and LBP-based features for texture image retrieval and classification. *Information Sciences*, *390*, 95–111. https://doi.org/10.1016/j.ins.2017.01.025

Mutch, J., & Lowe, D. G. (2006). Multiclass object recognition with sparse, localized features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *1*, 11–18. https://doi.org/10.1109/CVPR.2006.200

Naseer, A., Almujally, N. A., Alotaibi, S. S., Alazeb, A., & Park, J. (2024). Efficient Object Segmentation and Recognition Using Multi-Layer Perceptron Networks. *Computers, Materials and Continua*, *78*(1), 1381–1398. https://doi.org/10.32604/cmc.2023.042963

Naseer, A., Alzahrani, H. A., Almujally, N. A., Nowaiser, K. Al, Mudawi, N. Al, Algarni, A., & Park, J. (2024). Efficient Multi-Object Recognition Using GMM Segmentation Feature Fusion Approach. *IEEE Access*, *12*, 37165–37178. https://doi.org/10.1109/ACCESS.2024.3372190

Naseer, A., Mudawi, N. Al, Abdelhaq, M., Alonazi, M., Alazeb, A., Algarni, A., & Jalal, A. (2024). CNN-Based Object Detection via Segmentation Capabilities in Outdoor Natural Scenes. *IEEE Access*, *12*(June), 84984–85000. https://doi.org/10.1109/ACCESS.2024.3413848

Rafique, A. A., Gochoo, M., Jalal, A., & Kim, K. (2023). Maximum entropy scaled super pixels segmentation for multi-object detection and scene recognition via deep belief network. *Multimedia Tools and Applications*, *82*(9), 13401–13430. https://doi.org/10.1007/s11042-022-13717-y

Sikder, J., Islam, M. K., & Jahan, F. (2024). Object segmentation for image indexing in large database. *Journal of King Saud University - Computer and Information Sciences*, *36*(2), 101937. https://doi.org/10.1016/j.jksuci.2024.101937

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learning, ICML 2019*, *2019-June*, 10691–10700.

Telceken, M., & Kutlu, Y. (2022). Detecting Tagged People in Camera Images. *Journal of Intelligent Systems with Applications*, (May), 27–32. https://doi.org/10.54856/jiswa.202205197

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *Proceedings of Machine Learning Research*, *139*, 10347–10357.

Venugopal, V., Joseph, J., Vipin Das, M., & Kumar Nath, M. (2022). An EfficientNet-based modified sigmoid transform for enhancing dermatological macro-images of melanoma and nevi skin lesions. *Computer Methods and Programs in Biomedicine*, *222*, 106935. https://doi.org/10.1016/j.cmpb.2022.106935

Zhang, L., & Pu, J. (2024). Object recognition based on shape interest points descriptor. *Electronics Letters*, *60*(9), 1–3. https://doi.org/10.1049/ell2.13198

Zhang, R., Wang, L., Cheng, S., & Song, S. (2023). MLP-based classification of COVID-19 and skin diseases. *Expert Systems with Applications*, *228*(March), 120389. https://doi.org/10.1016/j.eswa.2023.120389