

## TRANSFER LEARNING FOR TURKISH CUISINE CLASSIFICATION

Sait ALP<sup>1\*</sup>


<sup>1</sup>Trabzon University, Faculty of Computer and Information Sciences, Department of Artificial Intelligence Engineering, 61335, Trabzon, Türkiye

**Abstract:** Thanks to developments in data-oriented domains like deep learning and big data, the integration of artificial intelligence with food category recognition has been a topic of interest for decades. The capacity of image classification to produce more precise outcomes in less time has made it a popular topic in computer vision. For the purpose of food categorization, three well-known CNN-based models—EfficientNetV2M, ResNet101, and VGG16—were fine-tuned in this research. Moreover, the pre-trained Vision Transformer (ViT) was used for feature extraction, followed by classification using a Random Forest (RF) algorithm. All the models were assessed on the TurkishFoods-15 dataset. It was found that the ViT and RF models were most effective in accurately capturing food images, with precision, recall, and F1-score values of 0.91, 0.86, and 0.88 respectively.

**Keywords:** Food classification, Deep learning, Convolutional neural network, Image classification, Transfer learning, ViT.

\*Corresponding author: Trabzon University, Faculty of Computer and Information Sciences, Department of Artificial Intelligence Engineering, 61335, Trabzon, Türkiye

E mail: saitalp@trabzon.edu.tr (A. ALP)

Sait ALP  <https://orcid.org/0000-0003-2462-6166>

Received: August 30, 2024

Accepted: October 28, 2024

Published: November 15, 2024

Cite as: Alp S. 2024. Transfer learning for Turkish cuisine classification. BJS Eng Sci, 7(6): 1302-1309.

### 1. Introduction

The problem of food recognition in still images has recently gained attention in the field of computer vision (Kiourt et al., 2020). While several benchmark datasets have been developed, featuring sample images of globally popular foods, a detailed analysis reveals a notable underrepresentation of Turkish cuisine. Despite the richness and diversity of Turkish dishes, they are scarcely represented in these datasets. A review of datasets created for food recognition indicates that there is only one dataset available that includes Turkish dishes. Consequently, the number of studies focusing on Turkish foods is insufficient, highlighting a clear need for more study to enhance the accuracy and applicability of food recognition systems for diverse culinary traditions.

The application of deep learning has significantly advanced the expanding field of computer vision, particularly in the context of image recognition tasks (Chai et al., 2021). A critical area of interest among these is food image recognition, which has emerged as a result of its extensive applications in health monitoring, dietary management, and interactive cooking guides (Chen et al., 2020; Zhang et al., 2023) Nevertheless, the complexity and diversity of food items, which are significantly different across various cultures and cuisines, present a challenge. This diversity requires image recognition systems that are scalable, versatile, and robust, and that can adjust to the variety of food presentation styles and appearances.

The initial studies on food recognition from images utilized manually created visual features and performed

food classification based on these attributes (Yang et al., 2010; Bossard et al., 2014; Beijbom et al., 2015). For example, Yang et al. (2010) proposed a representation that encodes the binary relationships of food ingredients based on local features and used Support Vector Machines (SVM) for classification. In another study, Bossard et al. (2014) introduced a classification approach based on identifying distinctive image parts using the Random Forest method. Beijbom et al. (2015) considered incorporating the menus of relevant restaurants by utilizing the location information of images to enhance recognition success.

Recent research has focused on leveraging the capabilities of Convolutional Neural Networks (CNNs) and domain adaptation techniques to enhance the accuracy and efficiency of food recognition systems. For instance, the study by Kayıkçı et al. (2019) demonstrates the application of CNNs in classifying Turkish cuisine on mobile platforms, highlighting the adaptability of deep learning models to function within the constraints of mobile devices while maintaining high accuracy. Meanwhile, Kawano and Yanai (2015) focused on expanding food image datasets through domain adaptation, integrating existing categories to create more comprehensive datasets. This approach not only creates culturally diverse food recognition systems but also reduces reliance on manual labeling through crowdsourcing. The study uses a "foodness" classifier and Adaptive SVMs to refine and expand food image datasets, enhancing their quality and applicability to a wider range of food categories, broadening the scope of food recognition systems.



Suddul and Seguin (2023) focuses on using AI, particularly computer vision and deep learning, for food type classification. The authors employed a dataset of over 16,000 food images spanning 11 categories and utilized data augmentation to address class imbalance. Three models were tested: CNN from scratch, transfer learning with InceptionV3, and transfer learning with EfficientNetV2. Among these, EfficientNetV2 achieved the best results, with a validation accuracy of 94.5% and an F1-score of 94.7%. Data augmentation, dropout, and early stopping techniques were applied to prevent overfitting. However, the dataset used in this study contains fewer food categories compared to the dataset I employed in my work.

Boyd and et al. (2024) investigates CNN architectures for fine-grained food image recognition, classifying 20 individual food items. The authors selected DenseNet after evaluating seven pre-trained models, achieving a baseline validation accuracy of 68%. Following parameter tuning, the optimized DenseNet model reached a validation accuracy of 79%. Notably, the dataset used in this study does not contain distinct food types but rather includes more specific food components, eggs, carrots, and butter, leading to significant visual differences between classes, which simplifies classification. In contrast, my dataset focuses solely on meals that exhibit considerable similarities, making the classification process more challenging.

As the demand for intelligent food recognition systems grows, the integration of advanced machine learning techniques will be crucial. The ongoing developments in this field promise to revolutionize how I interact with food through digital mediums, making technology an indispensable part of culinary experiences and dietary management.

CNNs excel in image processing tasks, making them a popular choice for object classification. Their architecture allows for the efficient categorization of hundreds of distinct classes. Building on the success of CNNs, transformers present a more recent advancement in deep learning. Initially designed for natural language processing, transformers have been adapted for image recognition. Their ability to handle sequential data and focus on relationships between different parts of the data makes them exceptionally effective. Unlike CNNs, which process data in a hierarchical manner, transformers process all parts of the data simultaneously, providing a more comprehensive understanding of the image content. This makes transformers particularly useful for complex image recognition tasks where context and relation between different image parts are crucial (Alp and Şenlik, 2023).

ViT is a pioneering approach that applies the principles of transformers, originally designed for natural language processing, to the domain of image recognition. Unlike traditional convolutional neural networks (CNNs) that process images through localized filters, ViT treats an image as a sequence of fixed-size patches, much like

words in a sentence. Each patch is embedded and then processed through a series of transformer blocks that utilize self-attention mechanisms.

This architecture enables the model to capture global dependencies between any parts of the image, which is beneficial for understanding complex scenes where contextual awareness is key. For example, in tasks like object detection or scene segmentation, ViT can leverage its global perspective to better differentiate and classify various elements within the image. Its ability to process all parts of the image simultaneously allows for a more comprehensive understanding of the entire scene, making it particularly useful for image recognition tasks that require a deep understanding of hierarchical and relational context.

Moreover, ViT has shown impressive performance on benchmarks, often surpassing traditional CNNs, especially when trained on large-scale datasets (Akan et al., 2023; Alp et al., 2024). This performance gain underscores the potential of transformer models to reshape the landscape of computer vision by providing a powerful alternative to established convolutional architectures.

Recent research has emphasized the use of ViTs and domain adaptation techniques to improve the accuracy and efficiency of food recognition systems.

Gao and et. al (2024) introduces AlsmViT, a Vision Transformer (ViT)-based method for food image classification, designed to handle visually similar foods. The model incorporates data augmentation (Augmentplus), deeper image processing (LayerScale), and enhanced feature extraction (MLP-GC). Tested on the Food-101 and Vireo Food-172 datasets, the AlsmViT-L model achieved validation accuracies of 95.17% and 94.29%, respectively. But its drawbacks include a relatively large number of model parameters, high computational demands, and significant peak memory usage.

Nijhawan et al. (2024) proposes a hybrid Vision Transformer (ViT) model for food cuisine detection, combining deep learning with hand-crafted features like GIST, HoG, and LBP. Using a dataset of 13 food categories, the model achieved 94.63% accuracy, 95.23% specificity, and 84.42% sensitivity, outperforming CNN-based models. By processing complete image data, it captures finer details, improving classification accuracy. However, the approach's computational cost is high due to the large number of tokens required.

While these models have an end-to-end architecture with a large number of trainable parameters, making the training process more costly, my proposed hybrid method utilizes a pretrained ViT network solely for deep feature extraction, resulting in no trainable parameters in that part of the model; the only trainable component is the classifier, which employs a Random Forest (RF) classifier, making the training process much faster and more cost-effective. These additions offer a more comprehensive view of the current research landscape

and position my contributions within the existing body of knowledge.

This study makes several important contributions to the field of food category recognition. Firstly, the study introduces a novel hybrid approach that combines a pre-trained ViT for feature extraction with a RF classifier. The classification performance of this hybrid method is better than that of traditional end-to-end CNN architecture. It also offers advantages such as efficient training, reduced computational costs, and faster training times, as it restricts the number of trainable parameters to only the classifier. Besides that, the study carefully compares how well the ViT-RF hybrid model works with three well-tuned CNN-based models, which are EfficientNetV2M, ResNet101, and VGG16. Lastly, this research is among the first to evaluate ViT-Based model on the TurkishFoods-15 dataset, contributing to the dataset's benchmark and providing a reliable comparison for future studies in the domain.

## 2. Materials and Methods

EfficientNetV2M, ResNet101, and VGG16 are deep convolutional neural network architectures that are frequently employed in image classification tasks. EfficientNetV2M (Tan and Le, 2021), the most recent member of the EfficientNet (Tan and Le, 2019) family, attempts to maintain a balance between the performance and the size of the model by scaling the network architecture. ResNet101, an extension of the ResNet (He et al., 2016) family, utilizes residual connections to resolve the vanishing gradient issue and facilitates the training of neural networks that are exceedingly deep. Conversely, VGG16 (Simonyan and Zisserman, 2015) is distinguished by its simplicity, which is characterized by the use of smaller convolutional filters and deeper network layers. In contrast to these CNN models, the ViT (Dosovitskiy et al., 2021) approaches image recognition

from a different angle. It leverages self-attention mechanisms typical of transformers used in natural language processing, treating image patches as sequences. This method allows ViT to focus on global dependencies between patches, making it highly effective for tasks requiring the recognition of complex patterns and details in large-scale images.

In the course of our investigation, I implemented fine-tuning on these three models. First, I eliminated the top layer from each model, as it was originally designed for the ImageNet dataset. I customized the models to correspond with the classification of food images by incorporating three additional layers. An AveragePooling2D layer was introduced as the initial step to perform spatial pooling and reduce the spatial dimensions of the features. A Flatten layer was then implemented to transform the pooled features into a vector representation. Subsequently, a Dense layer with ReLU activation was implemented to incorporate non-linearity and to capture intricate relationships within the data. The implementation of a dropout layer with a rate of 0.5 was necessary to prevent overfitting. In order to produce class probabilities for each food image, a Dense layer with SoftMax activation was incorporated as the final output layer (Table 1).

By adding more layers to the pre-trained networks, I was able to better adapt them to our particular food image classification task. Updating the weights of the new layers and freezing the weights of the previous layers was necessary to maintain the learned features during this process. I sought to optimize the classification performance of the models on our food images dataset by leveraging their pre-trained knowledge. Specifically, Table1 contains the parameters of the base-line and fine-tuned models, and the Table 2 contains the compile settings of the model respectively.

**Table 1.** Transfer learning models parameters

Base-Model	Total Parameters	Trainable Parameters	Non-trainable Parameters
EfficientNetV2M	53,174,723	21,775	53,152,948
ResNet101	43,186,575	528,399	42,658,176
VGG16	14,849,871	135,183	14,714,688
ViT-Base	86,389,248	-	-

**Table 2.** Model compile parameters

Parameter	Values
Image size	224×224
Batch size	128
Optimization	Adam (learning rate=0.001)
Loss function	Categorical Cross entropy
Epochs	100
Metrics	Categorical Accuracy
Call back	Save best Only

In addition to fine-tuning the convolutional neural network architectures, I also explored the utilization of the ViT for our classification tasks. The pretrained ViT model was used for extracting features, which encapsulate rich contextual and textural information crucial for distinguishing between various food items. These extracted features were then used as inputs for a RF classifier. The choice of RF was based on its ability to handle high-dimensional feature spaces and its strong classification performance, particularly when combined with ViT's extracted features. Its ensemble approach aggregates predictions from multiple decision trees, reducing overfitting and improving prediction accuracy. While alternative classifiers like SVM and KNN were considered, initial tests showed RF to be the best performer on this dataset.

This hybrid approach leverages the deep learning capabilities of ViT in feature extraction with the machine learning efficiency of Random Forest in classification, aiming to enhance the overall accuracy and reliability of the system.

For this study, I utilized the "TurkishFoods-15" (Güngör

et al., 2017) dataset, specifically developed to address the underrepresentation of Turkish dishes in existing food recognition datasets. Compiled by researchers from Hacettepe University, this benchmark dataset encompasses images of fifteen popular Turkish meals, with each class containing approximately 500 images. The dataset was primarily sourced from Google Images, following specific search queries to ensure relevance and variety.

The dataset was specifically selected for its comprehensive coverage of diverse food categories relevant to the study. Detailed information about the dataset, including the number of instances, categories, and features, is provided in Table 3.

The images underwent a rigorous cleaning process to eliminate irrelevant content, thereby ensuring the dataset's quality and applicability for training deep learning models. This dataset's creation aimed to provide a comprehensive resource for training and evaluating food recognition systems, particularly those aimed at recognizing Turkish cuisine.

**Table 3.** Overview of the Turkish foods-15 Dataset: number of instances, categories, and features

Food Name	# images in train	# images in test	# images
Biber dolmasi	436	48	485
Borek	686	76	762
Cig kofte	295	33	328
Enginar	411	46	457
Hamsi	303	34	337
Unkar begendi	283	32	315
icli_kofte	464	52	516
Ispanak	245	27	272
Kebap	784	87	871
Kisir	450	50	500
Kuru fasulye	434	48	482
Lokum	615	68	683
Manti	380	42	422
Simit	429	48	477
Yaprak sarma	454	50	504

### 3. Results and Discussions

Our experiments were conducted on a system equipped with an Intel(R) Core(TM) i5-7400 CPU, operating at 3.00 GHz, and paired with 8 GB of RAM. For graphical processing, the system was outfitted with an NVIDIA GeForce RTX 2080 GPU.

The food image classification task was used to assess the performance of our fine-tuned models, EfficientNetV2M, ResNet101, and VGG16, alongside ViT. The RGB image dataset, which comprises 15 classes, was utilized to train the models.

To assess the performance of the models, I divided the dataset into training and testing sets using an 80-20 split. This allocation ensures that 80% of the data is used for training the model, while the remaining 20% is reserved

for testing to evaluate model accuracy and generalization. Further, to fine-tune and validate our models during the training phase, 20% of the training set was set aside as a validation set. This validation subset allows for the adjustment of model parameters and helps prevent overfitting. This strategy ensures a comprehensive evaluation of the models' performance across unseen data, providing a robust measure of their predictive capabilities.

I removed the top layer from each model during training and added three additional layers: AveragePooling2D, Flatten, and two dense layers. I applied dropout regularization to mitigate overfitting. A categorical cross-entropy loss function was used to train the models, and the Adam optimizer was used to optimize them. I

evaluated all the models on TurkishFoods-15 dataset after training. The evaluation metrics were calculated for each class, as well as the macro average over all classes, and included precision, recall, and F1-score. Table 4 provides a comprehensive comparison of the performance metrics—precision, recall, and F1-score—across four deep learning models (EfficientNetV2M, ResNet101, VGG16, ViT) for the task of recognizing various Turkish dishes.

Vision Transformer (ViT) consistently demonstrates high precision, notably scoring the highest for several dishes like "Biber dolmasi" (0.96), "Enginar" (0.89), and "Kurufasulye" (0.96). EfficientNetV2M also performs well, particularly for "Hamsi" (0.97) and "Ispanak" (0.95). ViT again stands out with exceptional recall scores, especially for "Ispanak" (1.00) and "Simit" (0.98). ResNet101 shows strong recall for "Kisir" (0.98) and "Manti" (0.98). ViT achieves high F1-scores, excelling particularly with "Biber dolmasi" (0.92) and "Kurufasulye" (0.95). EfficientNetV2M and ResNet101 also show competitive F1-scores across several dishes, underscoring their balanced performance. ViT achieves the highest overall accuracy at 0.90, followed closely by EfficientNetV2M and ResNet101 both at 0.87. VGG16 lags slightly at 0.76. ViT leads in macro average F1-scores at 0.91, demonstrating its effectiveness across classes irrespective of class imbalance. In weighted average, ViT also tops the chart at 0.89, followed by EfficientNetV2M

and ResNet101 at 0.87.

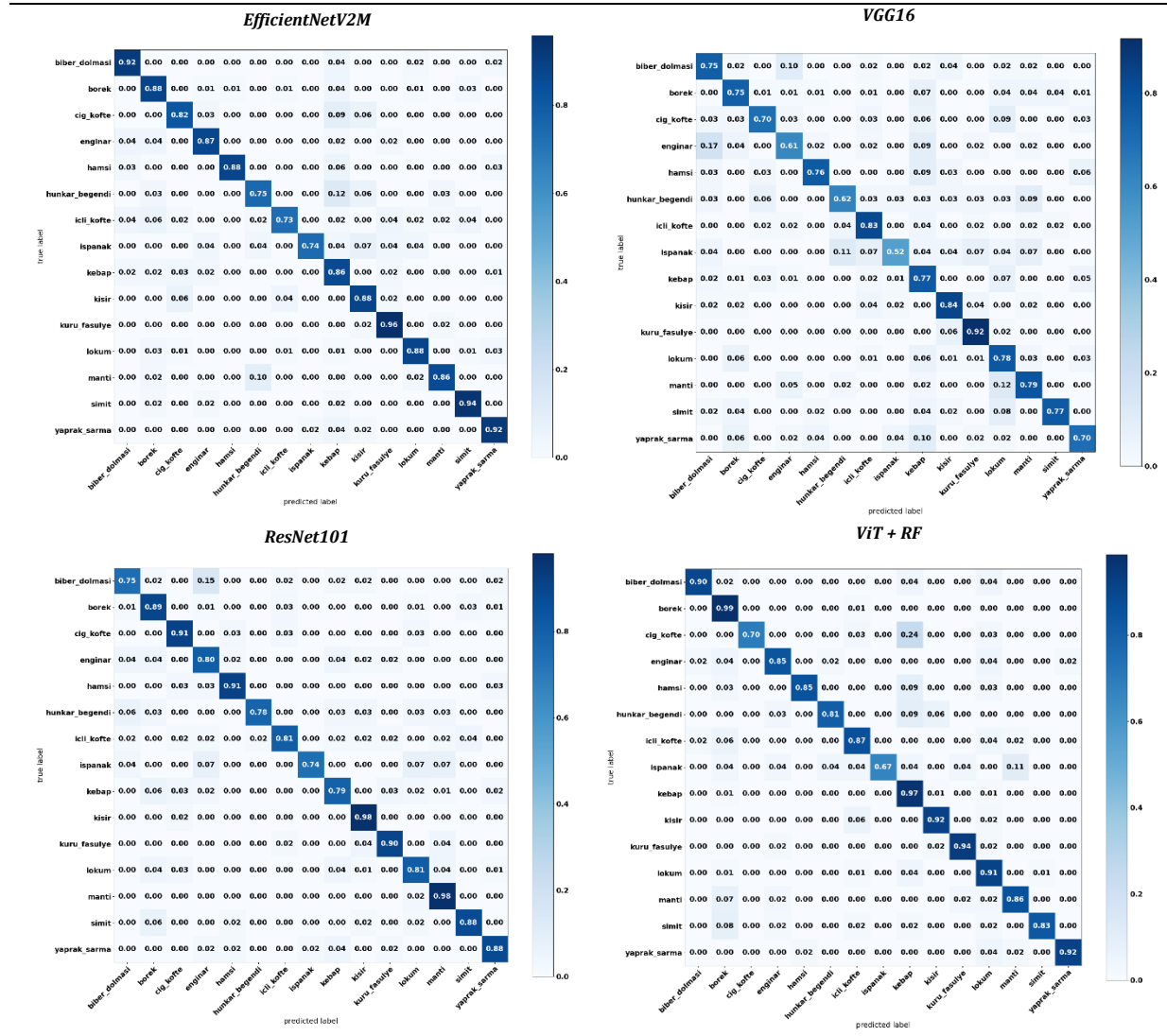
These results indicate that the Vision Transformer not only excels in individual categories but also maintains superior performance across the board, making it particularly effective for tasks requiring high precision and recall in image classification. EfficientNetV2M and ResNet101 also show robust performance, making them suitable alternatives depending on specific requirements like computational efficiency or model size. VGG16, while slightly less competitive, still offers reasonable accuracy for certain applications. Moreover, Table 5 show the confusion matrixes of all models.

The confusion matrix for EfficientNetV2M shows high diagonal values indicating strong class-specific accuracy for most dishes, with notable performance for "biber\_dolmasi" (0.92), "enginar" (0.87), and "kurufasulye" (0.88). These results suggest that EfficientNetV2M is quite effective in distinguishing between different Turkish foods, possibly due to its balanced scaling of depth, width, and resolution which enhances feature extraction across diverse image types. VGG16, known for its deep architecture and small convolutional filters, also performs well, particularly for "borek" (0.89) and "cig\_kofte" (0.91). However, it shows some confusion in classes like "enginar" and "hamsi," possibly due to the simpler and more uniform textures of these foods that challenge the model's deeper and narrower filters.

**Table 4.** The precision, recall, and F1-score for each class and the macro average were calculated

Food Name	Precision				Recall				f1-score				Support
	EfNetV2M	ResNet101	VGG16	ViT	EfNetV2M	ResNet101	VGG16	ViT	EfNetV2M	ResNet101	VGG16	ViT	48
Biber dolmasi	0.86	0.84	0.69	0.96	0.92	0.75	0.75	0.90	0.89	0.79	0.72	0.92	76
Borek	0.85	0.82	0.79	0.82	0.88	0.89	0.75	0.99	0.86	0.86	0.77	0.89	33
Cig kofte	0.77	0.79	0.74	1.00	0.82	0.91	0.70	0.70	0.79	0.85	0.72	0.82	46
Enginar	0.87	0.71	0.70	0.89	0.87	0.80	0.61	0.85	0.87	0.76	0.65	0.87	34
Hamsi	0.97	0.89	0.84	0.97	0.88	0.91	0.76	0.85	0.92	0.90	0.80	0.91	32
Unkar begendi	0.80	0.96	0.77	0.93	0.75	0.78	0.62	0.81	0.77	0.86	0.69	0.87	52
icli_kofte	0.90	0.89	0.78	0.85	0.73	0.81	0.83	0.87	0.81	0.85	0.80	0.86	27
Ispanak	0.95	0.95	0.74	1.00	0.74	0.74	0.52	0.67	0.83	0.83	0.61	0.80	87
Kebap	0.78	0.87	0.68	0.80	0.86	0.79	0.77	0.97	0.82	0.83	0.72	0.88	50
Kisir	0.85	0.86	0.81	0.92	0.88	0.98	0.84	0.92	0.86	0.92	0.82	0.92	48
Kuru fasulye	0.87	0.88	0.83	0.96	0.96	0.90	0.92	0.94	0.91	0.89	0.87	0.95	68
Lokum	0.92	0.86	0.67	0.81	0.88	0.81	0.78	0.91	0.90	0.83	0.72	0.86	42
Manti	0.92	0.80	0.70	0.88	0.86	0.98	0.79	0.86	0.89	0.88	0.74	0.87	48
Simit	0.90	0.91	0.90	0.98	0.94	0.88	0.77	0.83	0.92	0.89	0.83	0.90	50
Yaprak sarma	0.90	0.88	0.78	0.98	0.92	0.88	0.70	0.92	0.91	0.88	0.74	0.95	48
Accuracy	-	-	-	-	-	-	-	-	0.87	0.85	0.75	0.89	741
Macro avg	0.87	0.86	0.76	0.91	0.86	0.85	0.74	0.86	0.86	0.85	0.75	0.88	741
Weighted avg	0.87	0.86	0.76	0.90	0.87	0.85	0.75	0.89	0.87	0.85	0.75	0.89	741

Table 5. Models' confusion matrices



ResNet101 exhibits strong performance across several food classes with top accuracies for "kurufasulye" (0.92) and "manti" (0.98). The use of residual connections likely helps it maintain performance across deeper layers, improving the model's ability to learn from complex, varied food images. It also appears to handle intra-class variation effectively, likely due to its ability to leverage residual learning to avoid the vanishing gradient problem.

The Vision Transformer shows excellent performance, especially for "borek" (0.99) and "cig\_kofte" (0.91), demonstrating its capability to handle the relational context within food images effectively. Its architecture, which processes image patches through self-attention mechanisms, seems particularly adept at recognizing patterns and details critical for distinguishing similar food items.

Overall, the ViT and ResNet101 models show the most promising results, suggesting that architectures that can capture both long-range dependencies (ViT) and deep residual features (ResNet101) are beneficial for food image classification tasks involving complex and visually diverse dishes like those found in Turkish cuisine. EfficientNetV2M, while slightly less accurate in some classes, still performs robustly, suggesting its utility in scenarios where model scalability is crucial. VGG16, despite being an older model, holds up reasonably well, particularly in less complex food classes, highlighting its continued relevance in image classification tasks.

#### 4. Conclusion

This study examined the effectiveness of transfer learning approaches for categorizing Turkish cuisine images. I aimed to tackle the underrepresentation of Turkish dishes in existing food recognition datasets by fine-tuning CNN architectures, such as EfficientNetV2M, ResNet101, and VGG16, and using the pre-trained ViT for feature extraction and classification with a Random Forest algorithm.

Results showed that the combination of pre-trained ViT and RF outperforms CNNs in terms of accuracy precision, recall, and F1-score.

This combined method made good use of capacity of ViT to detect global dependencies in images, which made it adept at identifying Turkish cuisine's intricate patterns. Moreover, EfficientNetV2M and ResNet101 showed strong performance, making them good options for Turkish cuisine classification. Although VGG16 was not quite as good as the other models, it was still quite accurate, indicating that it is still useful for image classification tasks.

In particular, for diverse and culturally significant cuisines such as Turkish dishes, this study shows the potential of advanced deep learning and transformer-based models to improve food recognition systems. Expanding the dataset and looking into other deep learning architectures could be the focus of future research that aims to make classification more accurate

and faster in real-world situations.

One major limitation is that ViT models require large datasets and are computationally expensive, even though I only used them for feature extraction. This can make the approach less suitable for smaller datasets or low-resource environments. Additionally, while efficient, the RF classifier may not always generalize well to highly diverse datasets. In future work, I will explore using state-of-the-art ViT-based and CNN-based models for feature extraction, combined with multiple instance learning as the classifier, to improve performance and generalization. Testing on larger, more diverse datasets will also help enhance scalability and robustness.

#### Author Contributions

The percentage of the author contributions is presented below. The author reviewed and approved the final version of the manuscript.

	S.A.
C	100
D	100
S	100
DAI	100
L	100
W	100
CR	100
SR	100
PM	100
FA	100

C=Concept, D= design, S= supervision, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

#### Conflict of Interest

The author declared that there is no conflict of interest.

#### Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans. The authors confirm that the ethical policies of the journal, as noted on the journal's author guidelines page, have been adhered to.

#### References

- Akan T, Alp S, Bhuiyan MAN. 2023. Vision transformers and Bi-LSTM for Alzheimer's disease diagnosis from 3D MRI. The 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE), August 7-10, Las Vegas, NV, US, pp: 143.
- Alp S, Akan T, Bhuiyan MS, Disbrow EA, Conrad SA, Vanchiere JA, Kevil CG, Bhuiyan MA. 2024. Joint transformer architecture in brain 3D MRI classification: its application in Alzheimer's disease classification. *Sci Rep*, 14: 8996.
- Alp S, Şenlik R. 2023. Transfer learning approach for classification of beef meat regions with CNN. The 2023 Innovations in Intelligent Systems and Applications Conference (ASYU), August 14-16, Sivas, Turkiye, pp: 1-5.
- Beijbom O, Joshi N, Morris D, Saponas S, Khullar S. 2015. Menu-Match: restaurant-specific food logging from images. The 2015

- IEEE Winter Conference on Applications of Computer Vision, January 5-9, Waikoloa, HI, USA, pp: 844-851.
- Bossard L, Guillaumin M, Van Gool L. 2014. Food-101 – Mining discriminative components with random forests. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes Computer Sci, 8694: 446-461.
- Boyd L, Nnamoko N, Lopes R. 2024. Fine-grained food image recognition: A study on optimising convolutional neural networks for improved performance. *J Imaging*, 10(6): 126.
- Chai J, Zeng H, Li A, Ngai EW. 2021. Deep learning in computer vision: a critical review of emerging techniques and application scenarios. *Mach Learn Appl*, 6: 100134.
- Chen J, Zhu B, Ngo CW, Chua TS, Jiang YG. 2020. A study of multi-task and region-wise deep learning for food ingredient recognition. *IEEE Trans Image Process*, 30: 1514-1526.
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. 2021. An image is worth 16x16 words: transformers for image recognition at scale. URL=<https://arxiv.org/abs/2010.11929> (accessed date: August 31, 2024).
- Gao X, Xiao Z, Deng Z. 2024. High accuracy food image classification via vision transformer with data augmentation and feature augmentation. *J Food Eng*, 365: 111833.
- Güngör C, Baltacı F, Erdem A, Erdem E. 2017. Turkish cuisine: a benchmark dataset with Turkish meals for food recognition. The 2017 25th Signal Processing and Journal: Communications Applications Conference (SIU), May 15-17, Antalya, Türkiye, pp: 1-4.
- He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, Las Vegas, NV, USA, pp. 770-778.
- Kawano Y, Yanai K. 2015. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In: Proceedings of the Computer Vision - ECCV 2014 Workshops, September 6-, Zurich, Switzerland, pp: 3-17.
- Kayıkcı Ş, Başol Y, Dörter E. 2019. Classification of Turkish cuisine with deep learning on mobile platform. The 4th International Conference on Computer Science and Engineering (UBMK), September 19-21, Samsun, Türkiye, pp: 1-5.
- Kiourt C, Pavlidis G, Markantonatou S. 2020. Deep learning approaches in food recognition. In: Tshirintzis G, Jain L, editors. Machine learning paradigms. Learning and analytics in intelligent systems, vol 18. Springer, Cham, Germany, pp: 83-108.
- Nijhawan R, Sinha G, Batra A, Kumar M, Sharma H. 2024. VTnet+ handcrafted based approach for food cuisines classification. *Multimedia Tools Appl*, 83(4): 10695-10715.
- Simonyan K, Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition. URL=<https://arxiv.org/abs/1409.1556> (accessed date: August 31, 2024).
- Suddul G, Seguin JFL. 2023. A comparative study of deep learning methods for food classification with images. *Food Humanity*, 1: 800-808.
- Tan M, Le Q. 2019. EfficientNet: rethinking model scaling for convolutional neural networks. The 36th International Conference on Machine Learning, June 9-15, Long Beach, CA, US, pp: 6105-6114.
- Tan M, Le Q. 2021. EfficientNetV2: smaller models and faster training. The 38th International Conference on Machine Learning, July 18-24, Virtual Conference, pp: 10096-10106.
- Yang S, Chen M, Pomerleau D, Sukthankar R. 2010. Food recognition using statistics of pairwise local features. The 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, San Francisco, CA, US, pp: 2249-2256.
- Zhang Y, Deng L, Zhu H, Wang W, Ren Z, Zhou Q, Lu S, Sun S, Zhu Z, Gorriz JM. 2023. Deep learning in food category recognition. *Inf Fusion*, 98: 101859.