IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 November 2022, accepted 27 December 2022, date of publication 5 January 2023, date of current version 17 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3234519

APPLIED RESEARCH

MEFood: A Large-Scale Representative Benchmark of Quotidian Foods for the Middle East

MOHAMMED YUSUF ANSARI AND MARWA QARAQE
Division of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar
Corresponding author: Mohammed Yusuf Ansari (ma1@alumni.cmu.edu)

The Open Access funding is provided by the Qatar National Library.

ABSTRACT Automatic food recognition systems have been receiving increasing attention in the research community with the advancements in inductive learning (e.g., classification in computer vision) due to their applicability in the healthcare and hospitality industry. However, food recognition is challenging due to its fine-grained nature and its high correlation with culture, geo-location, and language. To make food recognition systems feasible for the Middle Eastern region, we present a large-scale dataset (MEFood) of commonly consumed food items in the Middle East, thereby providing a dataset for current development and establishing a benchmark for future research. We have also thoroughly examined the MEFood dataset highlighting its challenging aspects and its real-world nature. Additionally, we have conducted a thorough experimental study benchmarking the mainstream computer vision and mobile networks on classification, runtime, and resource utilization metrics. Our results highlight that EfficientNet-V2 achieves performance closer to the best-performing individual model on the MEFood dataset while having the least resource utilization and minimal inference times. Finally, we have performed a thorough error analysis study to glean additional insights about the networks and MEFood dataset.

INDEX TERMS Food recognition, benchmark dataset, computer vision, Middle Eastern cuisine.

I. INTRODUCTION

Food is an essential part of life; however, insufficient or excessive food consumption has negative consequences on the human body. In addition, certain diseases, such as Type I diabetes, requires the constant recording of all food intake to adjust insulin injection levels. This task is still a paper and pen based approach, which is inefficient and cumbersome. Similarly, food tracking is important for other scenarios, such as weight loss/gain and weight monitoring. There have been some applications developed that facilitate the recording of food intake, but these applications require a user to manually type and search for particular foods within a database in the application. In addition, with international travel and the rise of food rating applications, it has now become important to develop methods that can facilitate the automatic identification of foods in unfamiliar cuisines as well as streamline food identification and tagging in food rating applications. For instance, such systems are crucial for countries hosting large-scale sports events like the FIFA World Cup, Asian Games, and the Olympics to allow visitors the ability to seamlessly identify and explore different cuisines and dishes they may be unfamiliar with. Furthermore, food recognition and analytics also play an essential role in analyzing social media content by identifying the cuisine and dishes residents and tourists prefer. An extensive food analysis can provide crucial insights to the hospitality industry, thereby allowing for personalized food experiences for visitors and tourists from differing backgrounds. Another crucial application of AFR is quality assurance in large-scale kitchens, restaurants, and fast food chains. The AFR system can detect inadequate food preparation or presentation to warn the staff to re-prepare the food before serving.

Due to the aforementioned applications, food computing has been highlighted as an important research direction by the research community due to its benefits and wide use

The associate editor coordinating the review of this manuscript and approving it for publication was Shovan Barma.

cases. In recent years, automatic food recognition (AFR) has received renewed attention due to the success of deep learning models in classification tasks of computer vision and multimedia applications [1], [2], [3]. The benefits and application of AFR is wide and diverse. However, AFR is still challenging for many existing deep learning architectures because of its fine-grained nature. This is particularly true for cuisines with significant intra-class differences (i.e., diverse representation of a dish) and low inter-class variance (i.e., minimal visual differences between different foods). Middle Eastern cuisine, in particular, has significant intra-class differences and low inter-class variance. Furthermore, the visual representation of the same dish varies significantly between the different countries of the Middle East and North Africa (MENA) region. Subsequently, this non-rigid nature of food items makes it challenging for recognition models to extract relevant spatial features for effectively distinguishing between different food dishes [35], [40], [41].

In recent years, neural networks have become famous for image-based classification tasks because of their ability to extract relevant features, end-to-end characteristics, robustness, and state-of-the-art performance [7]. Bossard et al. [6] proposed a modification of the Alexnet architecture for the food recognition task and benchmarked its performance on the Food-101 dataset. Kyaga et al. [17] trained a custom CNN architecture and showcased that the deep learning outperformed the support vector machines and other conventional image recognition algorithms. Ao et al. [4] fine-tuned the GoogLeNet architecture to achieve the state-of-the-art accuracy in 2015. Similarly, Liu et al. [22] suggested the use of the fine-tuned deep convolutional neural network (DCNN) and transfer learning for advancing the classification accuracy on the Food-101 [6], UEC Food-256 [18], and UEC Food-100 [26] datasets. Myers et al. [27] introduced a system that recognizes food items in an image using semantic segmentation and predicts its nutritional content by utilizing volume estimation. Martinel et al. [25] proposed a custom architecture that combined slice convolutions and residual blocks to achieve state-of-the-art accuracy on the Food-101 dataset. Researchers have also introduced large-scale datasets (e.g., ISIA-500 [28], Food-2K [29]) to build food recognition systems with wide generalization capability. Jiang et al. [15] have proposed a deep learning framework that detects, classifies, and performs nutritional analysis on images with multiple food items. Recently, Qiu et al. [32] have introduced a deep learning framework (knowns as PAR-Net), which employs a convolutional neural network to generate global image classification, an auxiliary network to generate discriminative features, and a third network for extracting features for different image crops. The final prediction is based on the concatenation of the full image and the mined discriminative regions. The proposed framework achieves 90.4%, 90.2%, 92.0% accuracy on Food-101 [6], Vireo-172 [8], and Sushi-50 datasets, respectively.

The above-mentioned work use models trained with miscellaneous food datasets, thus limiting their extensive use in countries with a defined cuisine and food culture [5], [46]. To resolve this limitation, researchers have been developing datasets and tuning neural network architectures to cater to different cultures and cuisine. Subhi et al. [39] proposed a VGG-16 [38] based architecture to create a food recognition system for Malaysian cuisine. Tahir et al. [42] suggested a snapshot ensemble approach using MobileNet-V3 [14] to overcome sub-optimal local-minima convergence. Additionally, the authors present an explainable AI framework for food recognition and validate the performance of the proposed methodology on a comprehensive Malaysian food dataset. Similarly, Sahoo et al. [33] employed the ResNet and SENet architectures to create a large-scale food recognition system for southeast Asian cuisine with a special emphasis on Singaporean food. Jiang et al. [16] introduce a multi-scale multi-view feature aggregation scheme in neural network to utilize fine-grained ingredient information for food recognition. The authors validate the proposed framework on ChineseFood-Net [9] dataset. In a similar manner, Temdee et al. [44] suggested transfer learning and fine-tuning on the Inception-V3 model for Thai cuisine. One interesting application of AFR systems is highlighted by Sarker et al. [37], which identifies food items that may trigger hypertension and warns the user. To train robust food recognition networks, Mohanty et al. [30] have published a dataset containing 273 classes across 24,119 food images gathered from the real-world deployment of the MyFoodRepo app. One fundamental limitation of this dataset is the limited image count per class, which may hamper the performance of deep learning models. Recently, Qaraqe et al. [31] introduced a dataset for Middle Eastern dishes and utilized hand-crafted and deep features with particle swarm optimization and genetic algorithms to classify food popular in the Middle Eastern region. Although the dataset developed is unique and the first of its kind, it is limited in diversity and has insufficient images per class for training deep neural networks. In addition, the performance attained provides room for improvement.

To this end, the paper overcomes the shortcoming by providing the following contributions:

1) Builds a large-scale Middle Eastern food dataset (MEFood) that spans 70 classes (52,000 images) with approximately 744±125 images per class
2) Conducts a comprehensive experimental study that employs transfer learning and fine-tuning on a wide range of neural networks to benchmark the classification accuracy, parameter count, disk utilization, VRAM footprint, and training/inference times on the MEFood dataset
3) Performs an extensive analysis of the neural network's predictions to realize the inherent challenges of the proposed dataset and the trained models

The remainder of the paper is structured as follows: Section II describes the data crawling and cleaning

procedure and discusses the statistics of the MEFood dataset. Section III discusses the state-of-the-art neural network architecture included in the study for the AFR task on the MEFood dataset. Section V analyzes the performance of the benchmark AFR model and discusses the insights gleaned from the dataset and network predictions. Finally, Section VI summarizes the contributions of this work and discusses future directions.

II. DATA COLLECTION AND ANALYSIS

This section discusses the Middle Eastern food collection effort and analyzes the inherit similarities and differences between the different dishes and dishes within the same class.

A. FOOD CATEGORY DEVELOPMENT

The MEFood dataset aims to capture the diversity of food dishes in the Middle East with an additional emphasis on Gulf-based dishes. To develop a representative list of Middle Eastern and local based dishes, residents in Qatar were interviewed. The participants included Qatari nationals, people from the MENA region at large, and non-Arab individuals from around the world. The survey included questions to identify favorite traditional foods, comfort foods, most commonly consumed foods, and fast foods preferred in the Middle East region. The survey results highlight the diversification of the Qatari and Middle Eastern cuisine, with some dishes leading as favorites. The results were classified into four macro categories: namely, appetizers, main courses, desserts, and snacks. Fig 1 illustrates a sample of the images of different food dishes present in each macro category.

B. IMAGE COLLECTION AND CRAWLING

MEFood has been generated using multiple sources of images from the internet. Specifically, search engines such as Google and Bing were utilized to look up images of different food dishes. The top food images were downloaded using open-source crawlers.

It was observed that many of the images obtained from search engines contained high-quality food images intended for marketing purposes. To make the dataset representative and generalize with real-world plating and lighting, food images identified using different hashtags on the Instagram platform were also collected. It was found that the dataset generated by incorporating food images using different hashtags of Instagram was more representative of images that a user may snap using their mobile cameras. However, crawling images from Instagram was a challenge as the images resulting from a hashtag search resulted in many irrelevant results, compared to Google or Bing.

Another challenge faced in the collection and development of MEFood was that some Arabic dishes (e.g., Balaleet) did not result in enough images from the search engines with English queries. To overcome this barrier, Arabic queries were used to search dishes with low results in order to maximize the image count in the dataset. To avoid the duplication of images that are crawled from multiple sources, the following quality control measures in the post-crawling stage were taken. First, all the images were manually inspected and cropped to remove irrelevant images and to ensure that the ratio of food items to background is maximized. Second, duplicate food items were removed from the dataset by i) manual inspection and ii) running the images via a developed image similarity checker to identify images that were identical. Finally, to ensure that the dataset is balanced, an effort was put in collecting approximately 800 images per class.

Figure 1 and Figure 2 show a sample of the collected images under different macro-categories and illustrate the data distribution of the MEFood dataset, respectively. A total of 70 food/dish categories were identified, with an average number of images per class of 744 with a variance of 15625. Ultimately, MEFood presents a unique, representative, and multifarious Middle Eastern food dataset that accounts for multiple image sources, enabling its application in diverse scenarios.

C. MEFood DATASET ANALYSIS

Figure 3 highlights the complexity of the MEFood dataset. One of the intrinsic qualities and complexities of Middle Eastern type food is the high inter-class similarities. For example, many of the dishes compare similarly in terms of texture, color, and plating (i.e., Biryani, Maqlooba, and Mandi). This high inter-class similarity will serve as a challenge for the deep learning architecture to differentiate the different classes.

Another observation of the MEFood dataset is that there are significant variations within some of the same food classes, as depicted by Figure 3. A simple but clear example is pasta-based dishes. Due to different methods of preparation (e.g., type of pasta, base sauce, etc.), pasta-based dishes can have diverse shapes, colors, and textures, and appear visually different although the content is the same.

Collectively, MEFood contains food items widely consumed in the MENA region and aims to serve as a benchmark dataset that is challenging and representative for real-world use cases.

III. BENCHMARK MODELS

To validate the diversity and richness of the MEFood dataset, several wellknown benchmark models are trained and tested on the developed dataset. In specific, this section discusses the well know architectures that have been proposed in computer vision for image classification. We benchmark two mainstream models ResNet [13] and ConvNext [24] and two lightweight models MobileNet-V3 [14] and EffecientNet-V2 [43] on the MEFood dataset to establish a baseline for future research in AFR. A short description of the architectures of the models follows.

A. ResNet

He et al. [13] propose residual neural networks (ResNets) do overcome the problem of exploding and vanishing gradients in deep neural networks. The authors introduce residual

FIGURE 1. Four macro food categories with corresponding popular dishes, highlighting the diversity of MEFood dataset.

blocks, which employ skip connections for propagating the input information to the output, thereby improving information flow. As a result, the ResNets can have a depth of 152 layers (8xVGG depth). Limitations:

1) Complex architecture design and backpropagation strategy due to skip-connections.
2) Deeper networks require larger training datasets and longer training cycles.

B. MobileNet-V3

Howard et al. [14] suggest MobileNet-V3 architecture for maximizing performance and minimizing inference on machines with lower computational resources (i.e., computers with only CPUs and mobile devices). To minimize the computational resources required by the network, the authors employ a hardware-aware neural architecture search (NAS) (implemented by the NetAdapt [47]), which is supplemented with novel architecture design choices. Specifically, the network introduces squeeze and excitation blocks over the residual connections and a hard-swish non-linearity for enhancing performance on mobile devices. Limitations:

1) The paper introduces a method that modifies the results of NAS with an intuitive network design. However, the article doesn't suggest ways to accommodate intuitive network design within existing NAS techniques.

C. EfficientNet-V2

Tan et al. [43] present the second generation of efficient architecture (EfficientNet-V2) that aims to minimize training time and network size while maximizing the classification accuracy of the networks. The authors employ NAS to generate a combination of MBConv and fused-MBConv blocks, which can effectively scale with lower training times. Additionally, progressive training with adaptive regularization rates assists in faster network training with large image resolutions. Limitations:

1) The paper lacks evaluation of the lighter models of the EfficientNet-V2 family compared to MobileNet-V3 in terms of accuracy and inference time.

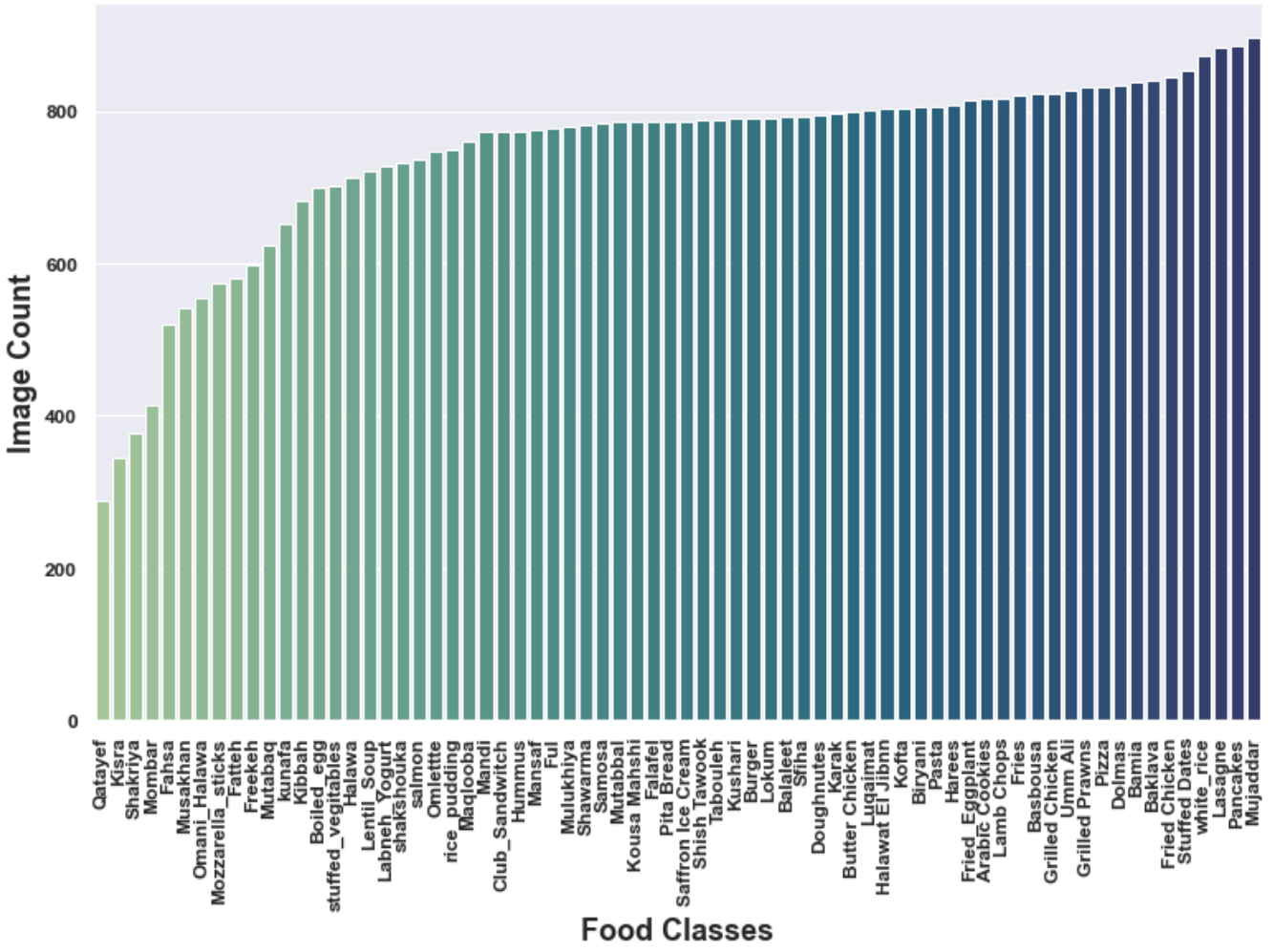

FIGURE 2. Global distribution of images in the MEFood, indicating overall balanced dataset.

FIGURE 3. Challenging images in MEFood dataset with low inter-class and high intra-class variability.

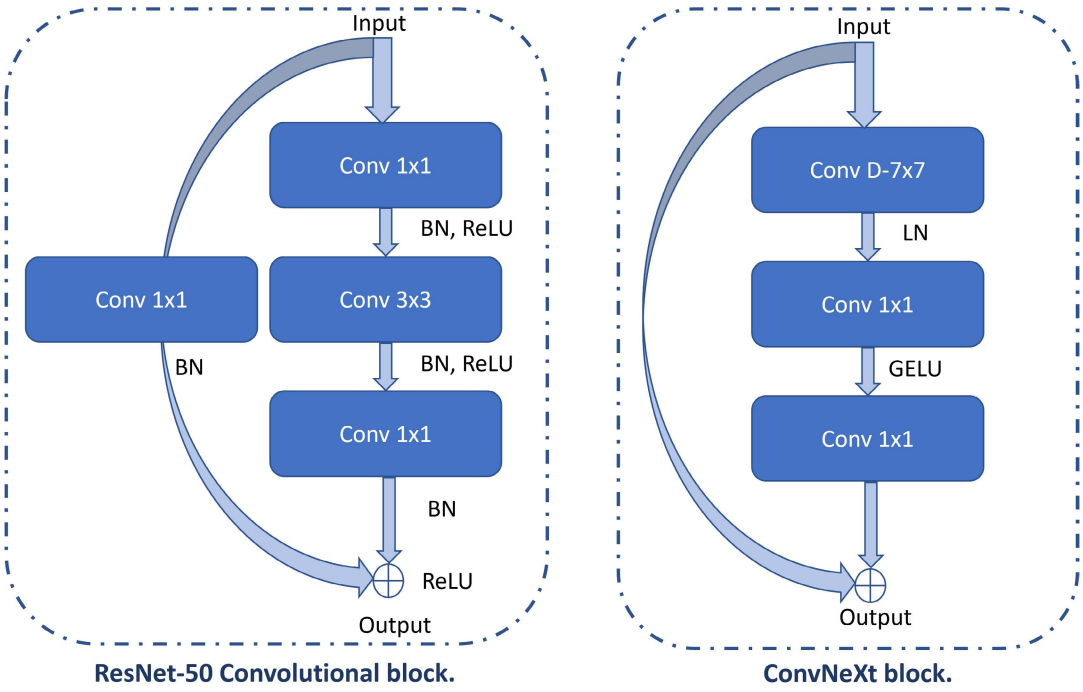

FIGURE 4. Convolutional building blocks of the ResNet-50 [13] and recently proposed ConvNext [24] architectures.

D. ConvNext

Liu et al. [24] propose the ConvNext architecture by improving the well-known residual neural network with modern training practices. ConvNext utilizes depthwise convolutions, layer normalization, and GELU activations to obtain higher accuracy. The network outperforms the recently proposed Swin transformer (i.e., vision transformer with convolutional priors) while maintaining the simplicity of the convolutional networks. Figure 4 highlights the innovative changes in the ConvNext architecture relative to the ResNet block. Limitations:

1) The robustness of ConvNext compared to the transformers has not been evaluated.
2) ConvNext may work best for specific tasks, while the transformers are more flexible and have applicability in tasks requiring discretized, sparse, or structured outputs.

IV. IMPLEMENTATION AND EVALUATION DETAILS

We implement the models benchmarked, mentioned previously, using Keras[1] framework. All images have been resized to 224 × 224×3 for all models except for ResNet-50, which requires input images of dimensions 229 × 229×3. We initialize all the models with Imagenet pre-trained weights. During training, we unfreeze (i.e., set as trainable) the last stack of residual blocks in ResNet-50 and ConvNext blocks (i.e., three blocks with 768 channels) in ConvNext_tiny. Similarly, we unfreeze the last three convolutional blocks in MobileNet-V3 large (i.e., expand_block 12, 13, and 14) and EfficientNet-V2B0 (i.e., blocks 6, 7, and 8). We add custom top layers (MLP) to each of these networks, which comprises global average pooling (GAP) followed by three fully connected layers (with 1024, 512, and 70 neurons). To minimize overfitting, the dropout layer is used after GAP and the first two fully connected layers (with probabilities 0.3, 0.35, and 0.25). We have implemented an image generator that feeds images to the network after applying random data augmentation such as horizontal flip, rotation (by at most 25 degrees), horizontal and vertical translations (by at most 10%), and zoom (upto 15 %) to reduce over-fitting and overcome other challenges of the MEFood dataset. We employ the categorical cross-entropy loss function with Adam optimizer and batch size of 32 to update the weights of the network. Every network is trained for 50 epochs for effective convergence. Furthermore, we save the weights at the end of epochs that resulted in maximal test set classification accuracy. At the end of training, we load the best weights of the networks to robustly evaluate the networks using a suite of classification metrics. We also measure the disk utilization and inference time of the networks on GPU and CPU (on a workstation, gaming laptop, and standard laptop). To elaborate, we perform ten consecutive evaluations of the test set on the CPU and GPU and report the average inference times. Figure 5 shows the complete workflow of fine-tuning neural networks and performing inference on the trained networks.

The workstation used for experiments is an HP Z8 workstation equipped with an Intel®Xeon(R) Silver 4216 CPU with

[1]F. Chollet, "keras," https://github.com/fchollet/keras, 2015

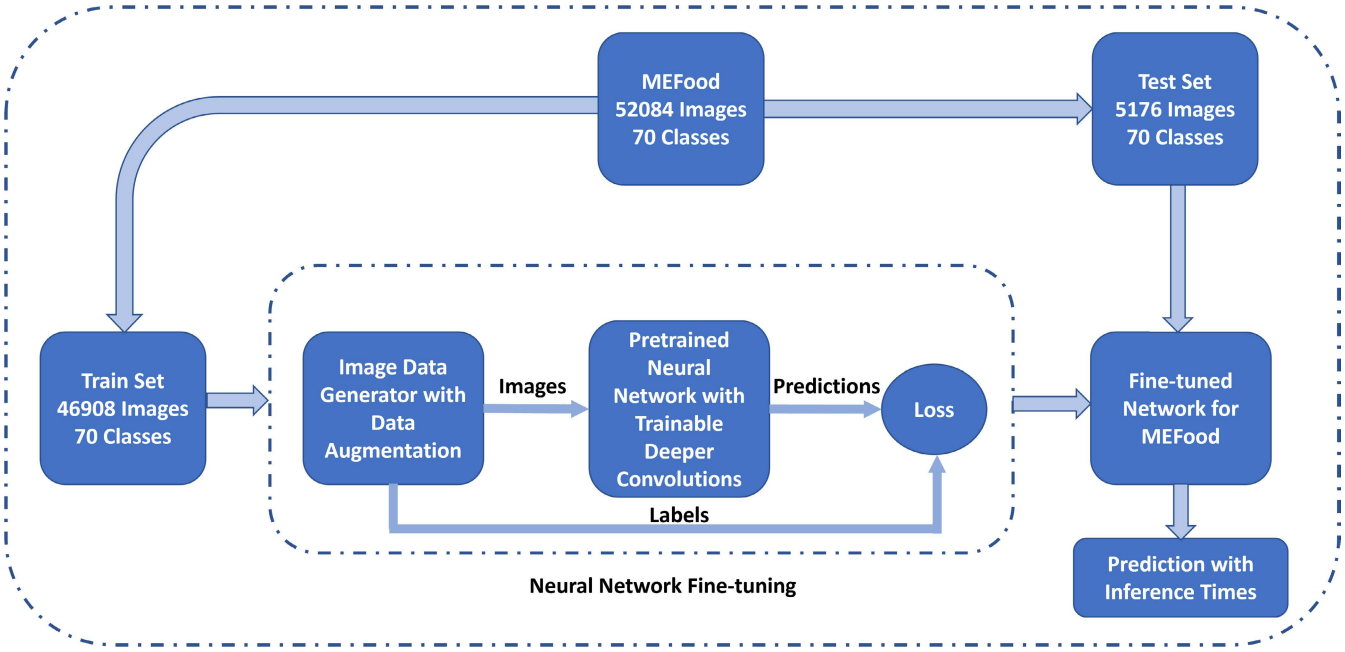

FIGURE 5. Flowchart showing the fine tuning and testing procedure on MEFood dataset.

a 2.10 GHz base clock (64 cores) and 64 GB of RAM. The networks were trained on a Nvidia Quadro RTX 5000 GPU with 16 GB of VRAM. For inference, we also employed a gaming laptop with a 6 core, 10th Gen Intel Core i7-10750H processor@2.6GHz (max turbo up to 5.0 GHz) and a standard laptop with a 6 core, 10th Gen Intel Core i7-10710U processor@1.1GHz (max turbo up to 4.7 GHz).

A. METRICS

In this subsection, we elaborate on the different classification metrics employed in our study to exhaustively evaluate the trained neural networks.

1) ACCURACY

Accuracy (also known as Top-1 accuracy) is computed by taking the ratio of correctly classified images to the total number of images in the evaluation data set. Top-5 accuracy is also computed by considering whether the correct label of the classified image is present in the Top-5 classes with the highest probabilities.

2) PRECISION

Precision computes the proportion of correctly classified positive images relative to all images classified as positive (i.e, *true positive/(true positive + false positive)*).

3) RECALL

Recall calculates the proportion of correctly classified positive images relative to actual positive images in the evaluation dataset (i.e, *true positive/(true positive + false negative)*).

TABLE 1. Classification accuracy of the bench marked neural networks.

Model Name	Acc (Top-1, Top-5)	Precision	Recall	F-1 score
ConvNext_tiny [24]	0.9466, 0.9901	0.9478	0.9467	0.9466
ResNet-50 [13]	**0.9493, 0.9934**	**0.9498**	**0.9494**	**0.9491**
EfficientNet-V2 [43]	0.9453, 0.9903	0.9457	0.9453	0.9450
MobileNet-V3 [14]	0.9422, 0.9895	0.9423	0.9422	0.9417

4) F-1 SCORE

F-1 score is a combined representation of precision and recall, obtained by calculating their harmonic mean.

V. PERFORMANCE ANALYSIS AND BENCHMARKING

In this section, we report and discuss the findings of our comprehensive empirical study aimed at evaluating the classification accuracy, disk utilization, GPU memory utilization, and training/inference speed of the benchmark models on the MEFood dataset.

A. CLASSIFICATION PERFORMANCE

Table 1 provides an exhaustive evaluation of classification metrics for four different benchmark models. Interestingly, we observe that ResNet-50 outperforms the ConvNext_tiny and the lightweight models on four different evaluation metrics. The higher performance of ResNet-50 over ConvNext_tiny, while having a lower parameter count suggests that higher classification accuracy on the ImageNet-1k dataset may not translate to downstream food recognition task. This is because food recognition is a **fine-grained** image classification, whereas mainstream computer vision emphasizes natural image classification. To elaborate, classifying food images is more challenging relative to natural images

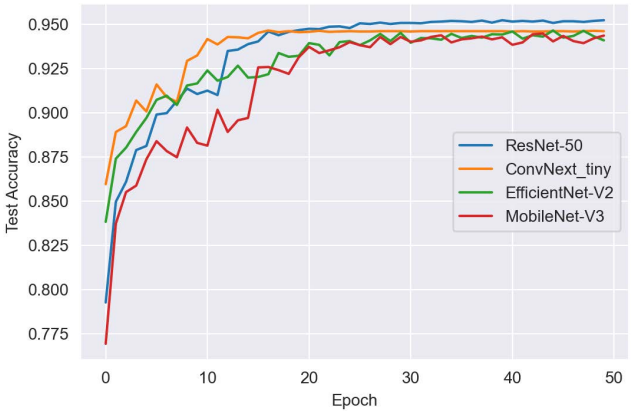

FIGURE 6. Plot showing the increase in Test set accuracy with epochs highlighting the learning ability of different neural networks.

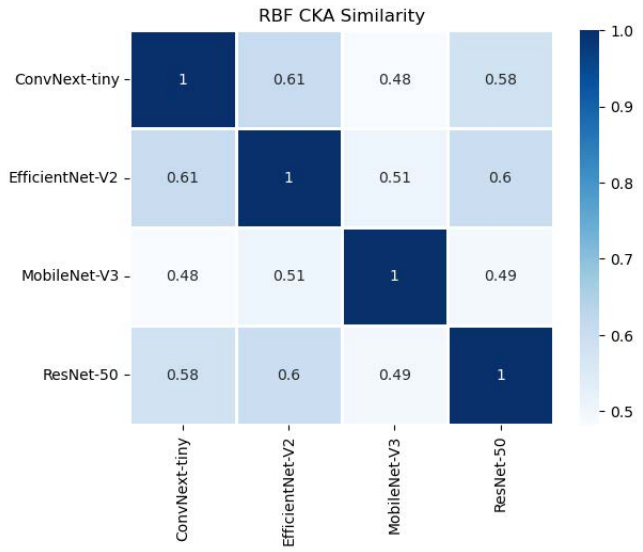

FIGURE 7. Confusion matrix showing the RBF CKA similarity score between the neural network representation generated by different neural networks.

due to additional factors, such as geolocation, culture, language, and changes in the quantity of ingredients. The variations in these additional factors may change the class of the food item (e.g., from Mandi to Khabsa). Therefore, it is crucial to propose network architectures or at the least network modules that can capture fine details in the image to *differentiate* food items with similar color and geometrical distributions. Nevertheless, we can infer from the table that all models attain high top-5 accuracy, implying that they can serve as a powerful logging and suggestion tool for applications that can utilize/process multiple system suggestions. We also observe that mainstream computer vision models that have been proposed for mobile devices (MobileNet-V3 Large), aiming to reduce training time and disk utilization (EfficientNet-V2) perform closely to the ResNet-50 and ConvNext, suggesting that larger networks with network logic for natural images may not have a performance advantage over lighter networks for fine-grained image classification tasks. Nevertheless, Figure 6 suggest that the larger networks with more parameters are able to generalize better in earlier epochs. Specifically, ConvNext_tiny achieves more than 90% test accuracy in less than 10 epochs. Among the lighter models, EfficientNet-V2 slightly outperforms MobileNet-V3 (Large) on all classification metrics. Additionally, Figure 6 indicates that EfficientNet-V2 generalizes better than MobileNet-V3 over the 50 epochs. We also observe that the precision, recall, and F1-score of each model lie closer to each other because the MEFood dataset has ample images per class for network training. Altogether, we have ensured that the MEFood dataset is a representative dataset for the Middle Eastern region. We encourage the research community to make innovative neural network modules to overcome the challenges in Arabian cuisine image classification tasks.

To interpret the feature representations generated by the different neural network architectures, we provide a similarity matrix generated using centered kernel alignment (CKA) [20]. Specifically, we use Radial basis function (RBF) CKA, which has been proposed to quantify the similarity between neural network representations generated from different parameter initialization. We can deduce from Figure 7 that ConvNext_tiny, ResNet-50, and EfficientNet-V2 generate similar feature encoding of food images (nearly 60% similarity). Interestingly, MobileNet-V3 feature representation has a lower similarity score with other network representations. This difference may be due to the constraints placed in the network architecture search (NAS) to minimize the computations on Mobile CPUs. On the other hand, EfficientNet-V2 NAS includes operations from the Fused-MBConv block, which is inspired by the residual block, thereby explaining the similarity in the representations between ResNet-50 and EfficientNet-V2.

TABLE 2. Disk utilization, training time, and GPU memory consumption of the bench marked neural networks.

Model Name	Parameter Count	Disk Utilization (MB)	Training Time (sec)	VRAM (MiB)
ConvNext_tiny [24]	29,166,758	117 MB	16650	3859
ResNet-50 [13]	26,248,634	105 MB	28500	11085
EfficientNet-V2 [43]	7,796,893	31.5 MB	10150	2695
MobileNet-V3 [14]	6,104,006	24.7 MB	10400	6791

TABLE 3. Inference time (in seconds) of the four benchmarked neural networks on the entire test set containing 5176 food images.

Model Name	Workstation GPU	Workstation CPU	Gaming Laptop CPU	Standard Laptop CPU
ConvNext_tiny [24]	107.22	254.68	434.67	632.78
ResNet-50 [13]	144.51	274.32	404.46	657.23
EfficientNet-V2 [43]	126.52	138.89	132.63	186.52
MobileNet-V3 [14]	107.88	134.90	123.24	170.26

B. PARAMETER UTILIZATION AND INFERENCE SPEED

Table 2 presents the parameter count, disk utilization, training time, and GPU memory utilization of the different models included in the empirical study. The selected models have a wide range of parameters to study the impact

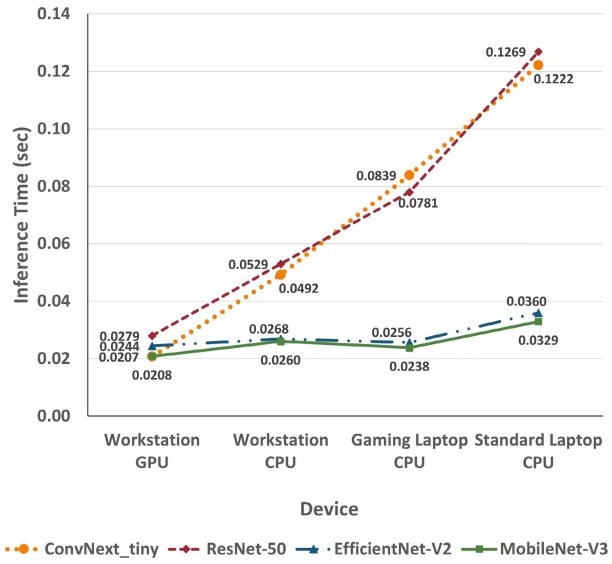

FIGURE 8. Plot describing the increase in inference time when moving from workstation to consumer hardware.

of parameter count on disk utilization and inference time. It can be observed that disk utilization is directly proportional to the parameter count of the models included in the study. MobileNet-V3 (large) has the lowest disk utilization of 24.7 MB, whereas the recently proposed ConvNext_tiny model has the highest utilization of 117 MB.

Interestingly, other factors like the training time and GPU memory consumption also have a significant impact on the usability and practical application of the neural networks. Shorter training times can allow the network to be re-trained for changing data distribution in the live environment. It is to be noted that the training times reported in Table 2 are for fine-tuning (i.e., retraining the deeper layers with a fully connected top) the networks to classify Middle Eastern foods. Even though ConvNext_tiny has more parameters than ResNet-50, it has significantly lesser training time (0.58×). The smaller training time of ConvNext_tiny is due to the macro and micro-architectural innovations within the model. The macro design choices include the use of patching strategy in the stem blocks, depthwise convolutions for spatial mixing, 1×1 convolution for channel mixing, inverted bottlenecks, and large kernel sizes. The micro changes involve using fewer activation and normalization layers, substituting batch-normalization with linear normalization, and applying the GELU activation function. These changes allow for ConvNext_tiny to have much lower FLOPS as compared to ResNet-50, thereby having lower training times. The design innovations in ConvNext_tiny also high a significant impact on the GPU memory (VRAM) consumption during inference, resulting in 0.35× VRAM utilization relative to the ResNet-50. Surprisingly, ConvNext_tiny attains lower VRAM consumption relative to MobileNet-V3 (large) model, suggesting that deep learning practitioners and researchers should consider VRAM utilization as an important factor while designing networks for mobile deployment. The faster training time and lower VRAM consumption of ConvNext_tiny with classification accuracy closer to ResNet-50 make it ideal for deployment in environments that require frequent retraining and have VRAM constraints (e.g., machines with low-end GPUs). We observe that lightweight models (i.e., EfficientNet-V2 and MobileNet-V3) have significantly less training time than ResNet-50 (0.36×) and ConvNext_tiny (0.62×). This is because the lightweight models have fewer parameters, requiring fewer computation and gradient calculations in forward and backward propagation, respectively. We find that EfficientNet-V2 has the lowest VRAM utilization among all the models in the empirical study, indicating that neural architecture search (NAS) should be employed with scaling strategies to optimize parameter count, VRAM utilization, and training efficiency.

We perform an inference study across GPU and CPUs of different machines to note the inference time variations under different deployment scenarios, thereby aiming to understand the performance of neural networks on low-end machines with limited computational resources. Table 3 presents the net inference time for the entire test set containing 5176 food images. Figure 8 visualizes the trend in inference times for predicting a single food image across different hardware. It can be observed from Table 3 and Figure 8 that the inference time of ConvNext_tiny and ResNet-50 models nearly doubles when shifting workstations GPU to workstation CPU. The increase in inference time is more for the ConveNext_tiny relative to ResNet-50, suggesting that operations within the ConvNext block are highly optimized for GPU-based computations. The inference times further increase linearly (Figure 8) when inferring on gaming and standard laptop CPUs, implying that these models may not scale well for low-end hardware with limited computational resources. On the other hand, EfficientNet-V2 and MobileNet-V3 experience a minor increase in inference time when shifting from workstation GPU to CPU. Surprisingly, the inference time for both models decreases when inference is performed on the gaming laptop CPU (fewer cores, higher frequency) instead of the workstation CPU (many cores, lower frequency). This suggests that lightweight models benefit from higher single-core gaming laptop performance. In other words, the core frequency of the CPU is more influential for the lightweight models relative to the core count. Altogether, the lighter models have a relatively minor increase in inference time across the different hardware testing environments. Among the tested models, we recommend the use of EfficientNet-V2 on resource-constrained machines because of its fast training and inference times, low disk and VRAM utilization, and performance that is closer to ConvNext_tiny and ResNet-50.

C. ERROR ANALYSIS AND ENSEMBLE APPROACH

We analyze the misclassified images of the four models in our study to understand the shortcomings and patterns across different neural network architectures. Specifically, we examine the food classes with more than eight misclassifications.

FIGURE 9. Visual analysis of misclassified food classes by the four different deep learning architectures showing similarities in color distribution and dish layout along with a count for misclassified images.

Figure 9 presents the food classes that are misclassified by multiple models along with the misclassification count. Kofta and lamb chops are misclassified by all four models. Similarly, Kofta and Falafel are misclassified by the lightweight models. This is because the circular Kofta shares the color and shape distributions with falafel, whereas Kofta on sticks appears similar to lamb chops. Some other common misclassification pairs by the models are Bamia as Ful, Mujaddar as Freekeh, Hummus as Mutabbal, Mutabbaq as Omelette, and Rice Pudding as Labneh. Upon closer observation, we can deduce that the misclassified images have similar textures and geometrical distributions in their macro ingredients (e.g., toppings and food containers). To elaborate, we can observe that Omlette, Mutabbaq and kofta, Falafel have been stacked on top of one another. Similarly, Rice Pudding, Labneh and Hummus, Mutabbal pairs share the plating layout (i.e., circular bowls). We believe that the similarity in food plating and colors in the images suggests the network that the two food images belong to the same food class, leading to misclassifications. Over the last two decades, several revolutionary modeling techniques (ranging from Bag of word (BoW) to convolutions) have been proposed in the literature to effectively capture and differentiate a wide range of textures common in the real-world [23]. However, it is challenging for neural networks to generate effective and robust representations of textures because of significant distortion, rotation, change of scale, illumination, and image degradation present in real-world datasets (e.g., MEFood). The computationally heavy nature of modern neural networks adds further complexity to generate compact representations for different textures on mobile and edge devices. Additionally, the texture classification benchmarks in the literature often employ high resolution images for network training, which differs significantly in resolution from food datasets (e.g., MEFood, ChineseFoodNet, UEC Food-100). Based on these limitations of the existing methodology, AFR needs efficient techniques for texture classification in low-resolution images to enable highly accurate while computationally efficient fine-grained food classification on mobile/edge devices.

Another consistent pattern across the misclassified classes of food is the similarity in color distribution, suggesting the trained neural networks rely heavily on the colors of the images to differentiate between the food items. Our finding is consistent with work presented by Christodoulidis et al. [10] for food recognition in dietary assessment. Surprisingly, the trained networks accurately classify the western food items and do not associate them with Middle Eastern food, indicating that there is a significant difference in the ingredients, color distribution, and food plating between the two cuisines. The supplementary material for different networks highlights the precision, recall, and F-1 score of each food class in the MEFood dataset, thereby validating our findings. Middle Eastern food is especially challenging to classify because most dishes have a similar base (e.g., Arabic bread or rice), protein (e.g., roasted chicken or lamb/beef), and spices. We believe that augmenting the food datasets with geo-location and macro food categories (e.g., appetizer, salad, main course, etc.) can greatly assist the neural network in identifying nearly identical dishes from different countries in the region, enabling the food logging and tracking systems to be personalized for each country rather than the whole region.

During the error analysis study, we observed that all models do not misclassify the same image because of their varying network architecture, which results in different decision boundaries. Based on this discovery, we implement a hard-voting ensemble approach to further improve the Top-1 accuracy of the models to 96.17% (i.e., 1.3% improvement in accuracy over the ResNet-50). However, the ensemble approach cannot be used for direct deployment on mobile devices due to limited computational capacity, memory, and disk constraints. Nevertheless, the increased performance of the ensemble approach and error analysis study suggests that the food computing community should aim to design efficient lightweight networks that can effectively capture texture, food plating layout, and geometrical distributions of objects (such as food containers and individual dish components). We believe that the attention mechanism [45] that is popularly used in vision transformers [11], [19] and modern convolutional networks can be tuned to identify fine-grain ingredients. However, conventional multi-head

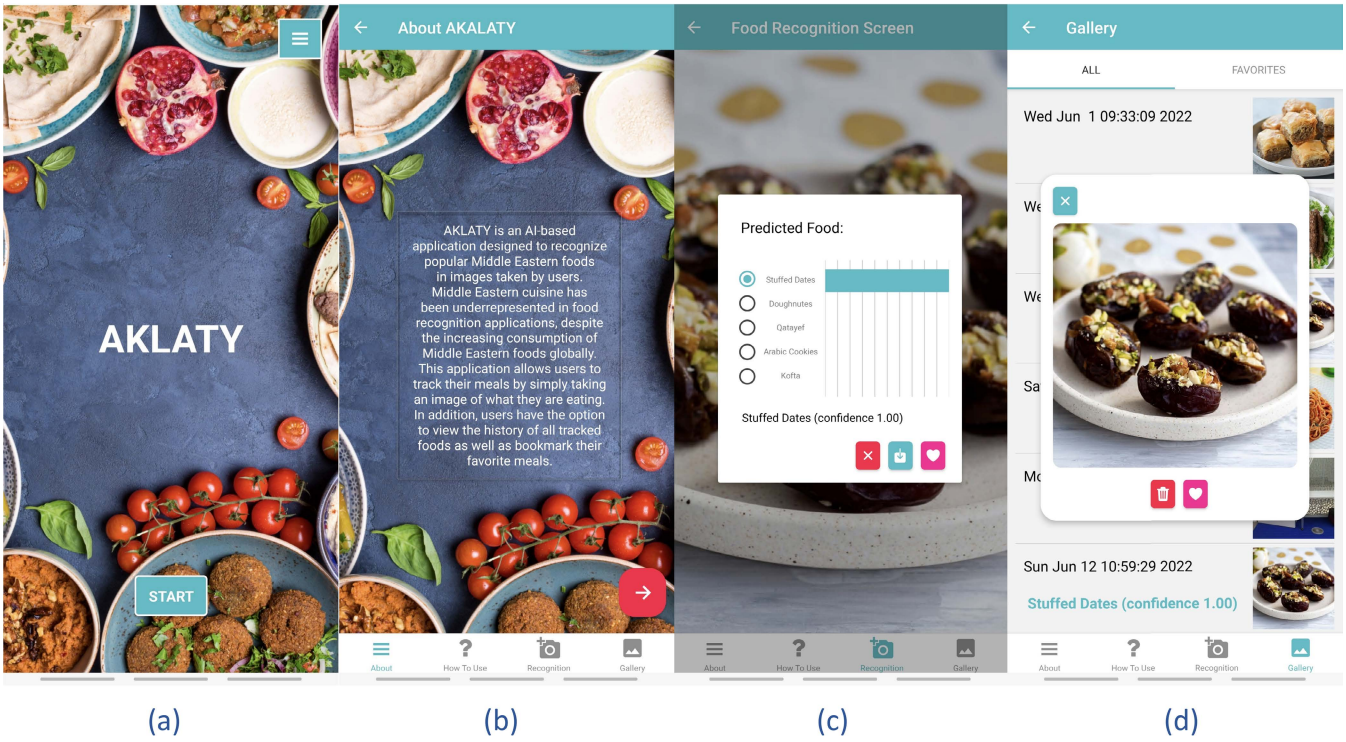

FIGURE 10. Different screens of Aklaty application, highlighting its goal and key features.

attention has quadratic computational complexity in terms of input size, requiring huge training datasets, computational power, and time. Therefore, lighter variants of attention mechanism having lower computational complexity should be adapted to enable real-time food recognition on devices with low computational power disk/memory constraints. A few lightweight variants of multi-head self-attention that overcome the quadratic computational cost have been proposed in mainstream computer vision [21], [34]. For instance, Guo et al. [12] present a self-adaptive linear attention mechanism that captures short and long-range correlations while avoiding the pitfalls of conventional attention mechanisms.

In the future, we aim to conduct a thorough experimental study of alternative linear time attention mechanisms for the food recognition task. We also aspire to design a network module based on linear cost attention to effectively capture the fine-grained intricacies of Middle Eastern food while providing real-time inference, low memory, and disk utilization. We believe that a revolutionary architecture for fine-grained image classification can allow researchers to tackle more challenging problems in healthcare (e.g., multi-cuisine and multi-regional food tracking) and the hospitality industry (e.g., food and advertising technology for restaurants).

D. UPCOMING AFR MOBILE APPLICATION

To accomplish our goal of developing an automatic food recognition and logging system for diverse use (diabetes care, hospitality, etc.), we have designed a minimal viable product (MVP) of our mobile AFR system based on the MEFood dataset. Figure 10(a) depicts the home screen of the mobile app "Aklaty"and (b) highlights the main aim of app. Figure 10 (c) is the food recognition screen that shows the Top-5 predicted classes (ranked by probability) for the food images captured through camera or uploaded from the mobile's gallery. Figure 10(d) shows the built-in gallery of recognized images that a user had previously logged. The app also provides user to bookmark their favorite food items. We aim to deploy Aklaty in the Middle Eastern region to help diabetes patients streamline their food tracking tasks while keeping dietitians/clinicians/doctors informed about their food habits.

Another target market for our app is teenagers and youth population the Middle East. Saudi Arabia has reported nearly 243 type-1 diabetes cases per 100,000 teenagers between the age of 13 and 16 [36]. As food imaging sharing is becoming popular on social media (e.g., Instagram and reels), we aim to capture the momentum to encourage teenagers to capture their daily food consumption using Aklaty. Next, Aklaty will perform recognition and provide healthier alternatives (i.e., rich in nutrition and low in calories) to encourage healthy eating habits. Aklaty will also provide informative summaries to teenagers about their overall food consumption. These summaries will highlight whether the total food consumed has calories lower/higher than the required amount (computed based on BMI) to show weight loss and weight gain patterns. As a whole, Aklaty aims to become a healthy food recommendation tool, which aims to make food recognition and logging seamless, while providing healthier food

recommendations and meaningful insights, while being tailored to commonly consumed foods in the Middle East.

VI. CONCLUSION

To conclude, we have assembled a first-of-its-kind, large-scale dataset of Middle Easter food images (MEFood) containing 70 different food classes with an average of 744 images per class. The dataset aims to capture the diversity in commonly consumed food items in Qatar and the Middle East region, providing ample images for data-hungry learning algorithms while serving as a benchmark for future research. We have analyzed and highlighted several challenges of MEFood. In addition, we have conducted a thorough empirical study benchmarking the performance of recently proposed mainstream computer vision networks and lightweight mobile networks. Specifically, we evaluated the networks on classification metrics, resource utilization, and inference times. Furthermore, we thoroughly analyzed the misclassifications of the networks to glean insights about their classification patterns. Based on our findings, we highlighted the key challenges in the fine-grained food classification and shortcomings of the existent neural network architectures. Finally, we presented some essential future directions to improve the state-of-the-art AFR and introduced our mobile application that aims to simplify diabetes healthcare.

ACKNOWLEDGMENT

The Open Access funding is provided by the Qatar National Library. The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] Y. Akhtar, S. P. Dakua, A. Abdalla, O. M. Aboumarzouk, M. Y. Ansari, J. Abinahed, M. S. M. Elakkad, and A. Al-Ansari, "Risk assessment of computer-aided diagnostic software for hepatic resection," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 6, no. 6, pp. 667–677, Jul. 2022.

[2] M. Y. Ansari, A. Abdalla, M. Y. Ansari, M. I. Ansari, B. Malluhi, S. Mohanty, S. Mishra, S. S. Singh, J. Abinahed, A. Al-Ansari, S. Balakrishnan, and S. P. Dakua, "Practical utility of liver segmentation methods in clinical surgeries and interventions," *BMC Med. Imag.*, vol. 22, no. 1, pp. 1–17, May 2022.

[3] M. Y. Ansari, Y. Yang, S. Balakrishnan, J. Abinahed, A. Al-Ansari, M. Warfa, O. Almokdad, A. Barah, A. Omer, A. V. Singh, P. K. Meher, J. Bhadra, O. Halabi, M. F. Azampour, N. Navab, T. Wendler, and S. P. Dakua, "A lightweight neural network with multiscale feature enhancement for liver CT segmentation," *Sci. Rep.*, vol. 12, no. 1, p. 14153, Aug. 2022, doi: 10.1038/s41598-022-16828-6.

[4] S. Ao and C. X. Ling, "Adapting new categories for food recognition with deep representation," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 1196–1203.

[5] B. Arslan, S. Memiş, E. B. Sönmez, and O. Z. Batur, "Fine-grained food classification methods on the UEC FOOD-100 database," *IEEE Trans. Artif. Intell.*, vol. 3, no. 2, pp. 238–243, Apr. 2022.

[6] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101—Mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 446–461.

[7] B. Chakravarthi, S.-C. Ng, M. R. Ezilarasan, and M.-F. Leung, "EEG-based emotion recognition using hybrid CNN and LSTM classification," *Frontiers Comput. Neurosci.*, vol. 16, pp. 1–9, Oct. 2022.

[8] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proc. MM*. New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 32–41, doi: 10.1145/2964284.2964315.

[9] X. Chen, Y. Zhu, H. Zhou, L. Diao, and D. Wang, "ChineseFoodNet: A large-scale image dataset for Chinese food recognition," 2017, *arXiv:1705.02743*.

[10] S. Christodoulidis, M. Anthimopoulos, and S. Mougiakakou, "Food recognition for dietary assessment using deep convolutional neural networks," in *Proc. Int. Conf. Image Anal. Process.* Springer, 2015, pp. 458–465.

[11] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6824–6835.

[12] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," 2022, *arXiv:2202.09741*.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[14] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[15] L. Jiang, B. Qiu, X. Liu, C. Huang, and K. Lin, "DeepFood: Food image analysis and dietary assessment via deep model," *IEEE Access*, vol. 8, pp. 47477–47489, 2020.

[16] S. Jiang, W. Min, L. Liu, and Z. Luo, "Multi-scale multi-view deep feature aggregation for food recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 265–276, 2020.

[17] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1085–1088.

[18] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 3–17.

[19] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10, pp. 1–41, Jan. 2022.

[20] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3519–3529.

[21] X. Li, Y. Jiang, M. Li, and S. Yin, "Lightweight attention convolutional neural network for retinal vessel image segmentation," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 1958–1967, Mar. 2021.

[22] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, and Y. Ma, "DeepFood: Deep learning-based food image recognition for computer-aided dietary assessment," in *Proc. Int. Conf. Smart Homes Health Telematics*. Springer, 2016, pp. 37–48.

[23] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikäinen, "From BoW to CNN: Two decades of texture representation for texture classification," *Int. J. Comput. Vis.*, vol. 127, no. 1, pp. 74–109, Jan. 2018.

[24] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11976–11986.

[25] N. Martinel, G. L. Foresti, and C. Micheloni, "Wide-slice residual networks for food recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 567–576.

[26] Y. Matsuda and K. Yanai, "Multiple-food recognition considering co-occurrence employing manifold ranking," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, 2012, pp. 2017–2020.

[27] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy, "Im2Calories: Towards an automated mobile vision food diary," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1233–1241.

[28] W. Min, L. Liu, Z. Wang, Z. Luo, X. Wei, X. Wei, and S. Jiang, "ISIA food-500: A dataset for large-scale food recognition via stacked global-local attention network," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 393–401.

[29] W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, and S. Jiang, "Large scale visual food recognition," 2021, *arXiv:2103.16107*.

[30] S. P. Mohanty, G. Singhal, E. A. Scuccimarra, D. Kebaili, H. Héritier, V. Boulanger, and M. Salathé, "The food recognition benchmark: Using deep learning to recognize food in images," *Frontiers Nutrition*, vol. 9, May 2022, Art. no. 875143.

[31] M. Qaraqe, M. Usman, K. Ahmad, A. Sohail, and A. Boyaci, "Automatic food recognition system for middle-eastern cuisines," *IET Image Process.*, vol. 14, no. 11, pp. 2469–2479, Sep. 2020.

[32] J. Qiu, F. P.-W. Lo, Y. Sun, S. Wang, and B. Lo, "Mining discriminative food regions for accurate food recognition," 2022, *arXiv:2207.03692*.

[33] D. Sahoo, W. Hao, S. Ke, W. Xiongwei, H. Le, P. Achananuparp, E.-P. Lim, and S. C. H. Hoi, "FoodAI: Food image recognition via deep learning for smart food logging," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2260–2268.
[34] R. Saini, N. K. Jha, B. Das, S. Mittal, and C. K. Mohan, "ULSAM: Ultra-lightweight subspace attention module for compact convolutional neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1627–1636.
[35] B. Sainz-De-Abajo, J. M. García-Alonso, J. J. Berrocal-Olmeda, S. Laso-Mangas, and I. D. L. Torre-Díez, "FoodScan: Food monitoring app by scanning the groceries receipts," *IEEE Access*, vol. 8, pp. 227915–227924, 2020.
[36] S. Saraswathi, S. Al-Khawaga, N. Elkum, and K. Hussain, "A systematic review of childhood diabetes research in the middle east region," *Frontiers Endocrinol.*, vol. 10, p. 805, Nov. 2019.
[37] P. Sarker, S. H. Islam, K. Akter, L. Rukhsara, and R. H. Hridoy, "A deep neural networks-based food recognition approach for hypertension triggering food," in *Proc. Int. Conf. Image Process. Capsule Netw.* Springer, 2022, pp. 360–373.
[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[39] M. A. Subhi and S. M. Ali, "A deep convolutional neural network for food detection and recognition," in *Proc. IEEE-EMBS Conf. Biomed. Eng. Sci. (IECBES)*, Dec. 2018, pp. 284–287.
[40] M. A. Subhi, S. H. Ali, and M. A. Mohammed, "Vision-based approaches for automatic food recognition and dietary assessment: A survey," *IEEE Access*, vol. 7, pp. 35370–35381, 2019.
[41] G. Tahir and C. K. Loo, "A comprehensive survey of image-based food recognition and volume estimation methods for dietary assessment," *Healthcare*, vol. 9, no. 12, p. 1676, 2021.
[42] G. A. Tahir and C. K. Loo, "Explainable deep learning ensemble for food image analysis on edge devices," *Comput. Biol. Med.*, vol. 139, Dec. 2021, Art. no. 104972.
[43] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.
[44] P. Temdee and S. Uttama, "Food recognition on smartphone using transfer learning of convolution neural network," in *Proc. Global Wireless Summit (GWS)*, Oct. 2017, pp. 132–135.
[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[46] L. Xiao, T. Lan, D. Xu, W. Gao, and C. Li, "A simplified CNNs visual perception learning network algorithm for foods recognition," *Comput. Electr. Eng.*, vol. 92, Jun. 2021, Art. no. 107152.
[47] T.-J. Yang, A. Howard, B. Chen, X. Zhang, A. Go, M. Sandler, V. Sze, and H. Adam, "NetAdapt: Platform-aware neural network adaptation for mobile applications," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 285–300.

MOHAMMED YUSUF ANSARI received the B.Sc. degree in computer science from Carnegie Mellon University and the M.Sc. degree in data science from Hamad Bin Khalifa University. He is currently pursuing the Ph.D. degree in computer engineering with Texas A&M University.

MARWA QARAQE received the M.Sc. and Ph.D. degrees in electrical engineering from Texas A&M University, College Station, TX, USA. She is an Associate Professor with Hamad Bin Khalifa University. Her research interests focused on machine learning and signal processing, including predictive analytics for health and security-related problems, such as developing robust predictive models for the early detection of diabetes, epileptic seizures, and stress and measuring attention in children with autism spectrum disorder via eye-tracking.