THE ASIAN BULLETIN OF BIG DATA MANAGMENT
Vol. 5. Issue 3 (2025)
https://doi.org/10.62019/kcsndc55

ASIAN BULLETIN OF BIG DATA MANAGEMENT
ISSN (Print): 2959-0795
http://abbdm.com/
ISSN (online): 2959-0809

Unlocking AI's Potential: Zero-Shot and Few-Shot Learning for Voice and Image Recognition

Adnan Ali, Shoaib Farooq, Muhammad Zeeshan Shafi, Muhammad Talha Tahir Bajwa*, Jamil Ur Rehman Hanifullah

Chronicle

Article history
Received: July 24, 2025
Received in the revised format: Aug 25, 2025
Accepted: Sept 13, 2025
Available online: Sept 27, 2025

Adnan Ali, is currently affiliated with the Department of Computer Science National College of Business Administration & Economics Lahore Sub Campus Bahawalpur, Pakistan.
Email: adnan.hnd@gmail.com

Shoaib Farooq is currently affiliated with the Department of Computer Science National University of Computer and Emerging Sciences (NUCES), Islamabad, Pakistan.
Email: m.shoaib1050@gmail.com

Muhammad Zeeshan Shafi is currently affiliated with the Department of Computer Science, the Islamia University of Bahwalpur, Pakistan.
Email: m.zeeshan.shafie@gmail.com

Muhammad Talha Tahir Bajwa is currently affiliated with the Department of Computer Science, University of Agriculture Faisalabad, Pakistan.
Email: talhabajwa6p@gmail.com

Jamil Ur Rehman is currently affiliated with the Group Head / Senior General Manager (IT) SSGCL, Department of IT, Pakistan.
Email: jamil94@yahoo.com

Hanifullah is currently affiliated with the Department of Computer science, Institute of business management sciences, Agriculture University peshawar, Pakistan.
Email: Hanifullahktk2003@gmail.com

Corresponding Author*

Abstract

The zero-shot and few-shot learning paradigms have emerged as promising solutions to the shortcomings of traditional deep models that require large volumes of labeled data for training. In the present paper, a full-fledged experimental study of the usage of zero-shot and few-shot learning methods in voice and image recognition tasks is provided. We analyze and compare several state-of-the-art architectures, such as CLIP, Whisper, and prototypical networks, in benchmark datasets including ESC-50 for audio classification and mini-ImageNet for image recognition. The experiments are designed to evaluate the generalization ability of models in situations where classes are unseen or sparsely represent during training. Our results show that multimodal models that are pre-trained on large-scale datasets have a high rate of performance in zero-shot scenarios, whereas metric-based few-shot approaches allow greater accuracy when only a small amount of supervision is provided. We also discuss cross-modal transfer ability, and examine how acquired representations in one modality (e.g., voice) can be used in the other (e.g., images). The findings highlight critical trade-offs between model complexity, data efficiency, and recognition accuracy which provide practical information to the deployment of lightweight and scalable AI systems in resource-limited settings. The paper develops the concept of generalized recognition systems and provides a base on how the concept will be researched on in the future in a low-resource learning environment. The main contributions of this work include a comparative study of ZSL and FSL methods, analysis of cross-modal transfer and identification of key trade-offs that lay the foundation for future research in generalized and adaptive AI.

Keywords: Zero-shot learning, Few-shot learning, Voice recognition, Image recognition, Cross-modal learning, Multimodal models, Low-resource AI, Deep learning, Transfer learning.

INTRODUCTION

The latest innovation in the field of deep learning has dramatically changed the picture concerning the recognition tasks, especially when it comes to computer vision and speech recognition. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) models have established state-of-the-art in numerous applications, including object recognition and speech synthesis. Nevertheless, these systems remain constrained by significant limitations in that they require large volumes of training data that are labeled. Gathering, cataloging and wrangling vast amounts of data is not only computationally expensive but also infeasible to most real-world applications, especially in low-resource settings. This difficulty is particularly acute in the situations when new categories or concepts are constantly being introduced and it is not possible to pre-label every potential example (Xian *et al.,* 2020). This has led to the development of the solution known as zero-shot learning (ZSL) and few-shot learning (FSL). The two paradigms target generalization of learning using limited or no labeled data through the use of auxiliary information, including semantic embeddings, attributes, or prior knowledge. This change in the data-intensive methods to the data-efficient methods of learning has been instrumental in areas where such annotated data is either scarce or costly to access.

DEFINITIONS AND BACKGROUND

Zero-Shot Learning (ZSL), and Few-Shot Learning (FSL) are specific paradigms of machine learning that seek to address the shortcomings of data-intensive deep learning models.

- **Zero-Shot Learning (ZSL):** ZSL is the capacity of one model to label objects or classes that it has never encountered in training with the assistance of semantic embedding information, attributes, or descriptions.
- **Few-Shot Learning (FSL):** FSL is the challenge of having a model generalize to new categories by actually having only a few labeled examples (typically 1-5 per class). This is usually done through meta-learning methods, metric learning or prototypical networks.

The difference between the zero-shot and few-shot learning in voice and image recognition is shown in Figure 1. Those paradigms have a close connection with transfer learning and meta-learning because both of the mentioned ones strive to apply the knowledge gained on previously observed tasks or data to unseen ones. The formalization of the ZSL and FSL definitions enable researchers to develop a base upon which systems can be designed to meet the dynamic and low-resource requirements.

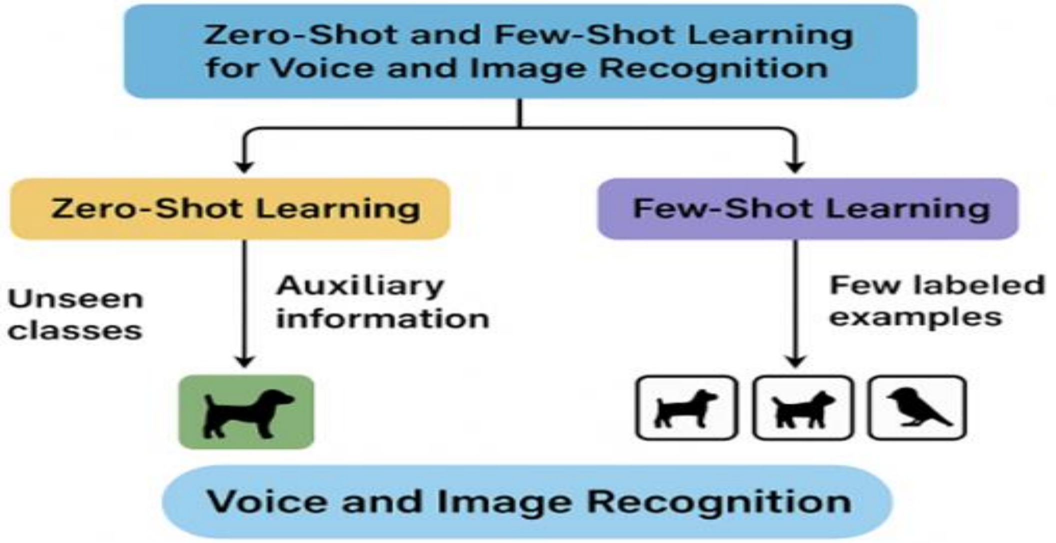
Zero-Shot and Few-Shot Learning for Voice and Image Recognition
Zero-Shot Learning
Few-Shot Learning
Unseen classes
Auxiliary information
Few labeled examples
Voice and Image Recognition

Figure 1.
Zero-Shot vs. Few-Shot learning in voice and image recognition

Zero-Shot and Few-Shot Learning: Concepts and Motivation

Both zero-shot learning (ZSL) and few-shot learning (FSL) aim to solve the issue of the scarcity of labeled data, with these systems being able to identify or classify examples not presented in their training set.

Zero Shot Learning (ZSL)

Zero-shot learning enables models to be made to predict unseen classes by utilizing auxiliary information, including semantic embeddings or attribute descriptions. This method does not require a labelled data of every new class, but instead it uses knowledge that can be shared across classes. For example, a ZSL model based on animal classes might identify a zebra without having ever been shown an instance of one in training, by relying on descriptions such as horse-like with black and white stripes or Equid with unique markings. In projecting objects onto a semantic space and correlating them with observed classes, ZSL is able to recognize entirely novel objects.

Few Shot Learning (FSL)

Few-shot learning (FSL) on the other hand allows models to transfer to new classes through training with a small number of labeled examples, which can be as few as one or five examples. FSL is most commonly accomplished by means of such approaches as meta-learning (e.g., Model-Agnostic Meta-Learning or MAML), prototypical networks, and exemplar-based learning. The techniques are aimed at ensuring that the model learns to generalize using few examples (Chen *et al.,* 2025). The examples would be a few-shot learning model that is trained to learn new voice commands or categorize a new set of animals using few labelled instances at each class.

Motivations for ZSL and FSL

The rationale behind both ZSL and FSL is simple: in practical usage we frequently have rare categories that either are costly to label or are hard to elicit in scale. The examples are voice identities in speaker recognition tasks, or fine-grained image categories in visual recognition tasks, where one may not be able to accumulate sufficient data. ZSL and FSL provide an opportunity to design scalable, versatile systems that do not heavily depend on large training data by allowing models to identify or categorize objects based on limited or no labeled information (Jogi *et al.,* 2025). This can be very handy when there are constraints in resources like the mobile devices, where data storage and computations are minimal.

APPLICATIONS TO VOICE AND IMAGE RECOGNITION

Image Recognition

ZSL and FSL have seen a remarkable forward in image recognition by the creation of multimodal models. An example of such is CLIP (Contrastive LanguageImage Pretraining), a model which is trained to embed both text and visual representation into a common semantic space. With the aid of image-text-image matching, CLIP allows zero-shot classification: the model can be used to classify images of categories they have never seen based on textual descriptions of those categories (Radford *et al.,* 2021). As an example, a query of a photo of a Dalmatian dog gives CLIP the ability to access pertinent images among a collection of hidden images, which might not have been a part of its training set. Such techniques as CoOp (Context Optimization) and CuPL (Contrastive Prompt Learning) have been demonstrated to perform fine-tuning on models such as CLIP using limited quantities of task-specific data in the few-shot setting. The methods enable the model to be trained on new tasks fast by changing a limited number of parameters on the basis of the limited examples. It has also been demonstrated that prompt tuning (as in CoOp) greatly increases the capacity of the model to learn new categories using a limited amount of data.

Voice Recognition

The voice recognition with ZSL and FSL has provided new opportunities in real time and cross-domain tasks. An example approach is voiceprompter, a zero-shot voice converter that relies on in-context learning and conditional flow matching to adjust to new, unseen speakers. It has proven to be much more similar in speakers and natural even when the speaker in question has no previous data (Choi and Park, 2025). The potential uses of this approach are enormous in security and personalization where voice recognition when using minimally trained data is a major challenge (Yin et al., 2025; Qadiret al.,2024).

An additional future development is RT-VS (Real-Time Voice Synthesis) which applies differentiable speech processing and articulatory characteristics to perform real-time zero-shot voice conversion. This enables speech of one speaker to be translated into another without the need to have huge datasets of comparable speech samples (Liu *et al.,* 2025). Equally, the multilingual automatic speech recognition (ASR) system known as Whisper has demonstrated a remarkable few-shot performance on underrepresented languages, an attribution that demonstrates its ability to recognize speech despite having a small in-domain sample (Sehar *et al.,* 2025; Li et al.,2024; Huo et al., 2024).

Cross-Modal and Multimodal Synergies

The cross-modal learning can integrate strong synergies between various forms of data, including images and audio. Cross-modal learning is based on the fact that a model can take the knowledge gained in a particular area (e.g., visual data) to assist in the recognition task in a different area (e.g., audio). This method has been of particular help in biometric matching, whereby a voice print is applied to either verify or identify the faces or vice versa. Indicatively, multimodal embeddings such as ImageBind bring audio, image, and text inputs into one model, and they are capable of zero-shot and few-shot learning tasks entailing multiple modalities (Girdhar *et al.,* 2023).

The Figure 2 demonstrates that cross-modal transfer allows knowledge acquired in one mode, e.g., voice, to facilitate the recognition in another mode, e.g., images. This may involve activities such as audio-to-image retrieval, in which an audio sample (e.g., of a dog barking) may be used to find appropriate images of dogs. Cross-modal models are also flexible in nature and can be applied to task types such as speech emotion recognition, whereby speech characteristics can be used to predict image-based mood or the opposite (Su *et al.,* 2025; Abaidullah & Basheer, 2024). Additionally, transfer learning across data of varying types can be enhanced by multimodal learning to make use of fewer large labeled datasets per modality (Basheer et al., 2024).

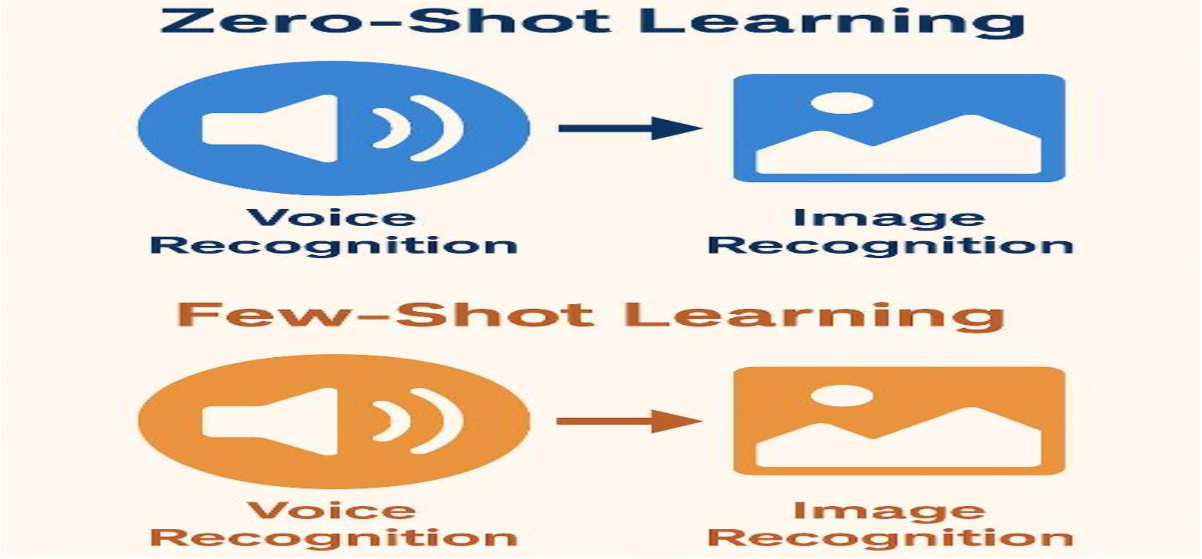

Figure 2.
Cross-modal transfer between voice and image recognition

Biometric Matching

Previous studies on the use of cross-modal biometric matching (e.g., voice to face matching) have shown how well audio cues can be used to boost visual recognition. Using a combined representation of voice and visual dimensions, these systems can achieve higher accuracy in any given task, such as recognizing individuals by voice and appearance, with only limited training information on one of the modalities (Nagrani *et al.,* 2018). The cross-modal opportunity presents numerous useful applications, including more secure authentication or enhancing human-robot interface through visual and auditory stimulus.

LITERATURE REVIEW

Early ZSL techniques match their visual appearance to a semantic space (attributes or word embeddings) such that models can make judgments of unseen classes. DeViSE mapped images into a text-learned word-embedding space (Frome *et al.,* 2013), ALE and ESZSL learned bilinear compatibility and simple linear embedding as strong baselines (Akata *et al.,* 2013; Romera-Paredes and Torr, 2015). The generalized ZSL work revealed the bias in seen-class classification and determined effective evaluation procedures (Xian *et al.,* 2018). Attributes annotations and fine-grained datasets (e.g., CUB) drive advances by emphasising semantic transfer (Wah *et al.,* 2011). FSL is dominated by metric-based and meta-learning. A similarity function was learned by siamese networks in one-shot recognition (Koch et al., 2015). Non-parametric nearest-prototype inference was proposed in an embedding space by Matching Networks (Vinyals *et al.,* 2016) and Prototypical Networks (Snell *et al.,* 2017). Relation Networks was trained on a learnable similarity metric (Sung *et al.,* 2018), whereas MAML offered a gradient-based meta-learner to learn quickly (Finn *et al.,* 2017).

An extensive re-assessment highlighted such concerns as overfitting to benchmark splits and the significance of more powerful baselines (Chen *et al.,* 2019). Image-text zero-shot recognition has become a de facto by contrastive pretraining on image-text pairs. In addition to CLIP, the ALIGN and LiT scaled data and enhanced the initialisation to robust zero-shot transfer (Jia *et al.,* 2021; Zhai *et al.,* 2022). BLIP-2 separates the encoders of the vision off of the LLM using lightweight Q-Formers to allow sample-efficient adaptation (Li *et al.,* 2023). To optimize few-shot transfer in particular, prompt-learning and adapters are useful: CoOp/CoCoUp can optimize prompts to new domains (Zhou *et al.,* 2022a, 2022b), and Tip-Adapter can use an additional non-parametric cache, which is activated on CLIP features with strong few-shot gains (Zhang *et al.,* 2022). Speech models that are self-supervised, such as wav2vec 2.0 and HuBERT, are trained on universal acoustic representations transferring to low-resource ASR and classification (Hsu *et al.,* 2021). SpeechT5 cross-modal transfer of speech and text (Ao *et al.,* 2022). To have a strong zero/few-shot ASR and command recognition, such representations have been tested on the datasets including Speech Commands and ESC-50 (Warden, 2018; Piczak, 2015). In zero-shot voice conversion/TTS, AutoVC, StarGAN-VC, and YourTTS, speaker disentanglement and generalization across speakers was exhibited with minimal target data (Qian *et al.,* 2019; Kameoka *et al.,* 2018; Casanova *et al.,* 2022).

In cases of lack or scarcity of one of the modalities, audio-visual pretraining improves recognition. AV-HuBERT is trained simultaneously with audio and lip movement, enhancing the resilience to the situation with low resources (Shi *et al.,* 2022). Large-scale audio-visual datasets (e.g., VGGSound) and common embedding spaces are found to be advantageous to cross-modal retrieval and grounding and allow zero-shot transfer of modalities (Chen *et al.,* 2020). ZSL centrally relies on few-shot image benchmarks, including miniImageNet and tieredImageNet (Vinyals *et al.,* 2016; Ren *et al.,* 2018), whereas attribute-rich and fine-grained datasets (CUB, AwA2) are key to it (Wah *et al.,* 2011; Xian *et al.,* 2017). In audio, there are ESC-50 (environmental sound classification), Speech Commands (keyword spotting) (Piczak, 2015; Warden, 2018). The recent surveys focus on Standardized splits (particularly generalized ZSL), episodic assessment when using FSL, and domain shift tests to capture actual deployments (Xian *et al.,* 2018).

Zero-shot recognition Multimodal encoders trained in a given modality are consistently superior in modalities, whereas metric/meta-learning performs well when a small number of labeled examples are provided. Prompting/adapters provide a low-weight interface between the two: they maintain zero-shot generality but do allow rapid few-shot specialization (Zhou *et al.,* 2022a; Zhang *et al.,* 2022). Self-supervised encoders produce less labeled data in speech, and might need domain adaption to accented, noisy, or low-resource languages (Baevski *et al.,* 2020).

METHODOLOGY

Research Design

The research design in this case is an experimental research design which will evaluate the effectiveness of the zero-shot and few-shot learning technique in the voice as well as image recognition tasks. The experiments will be used to test the capacity of the state-of-the-art models to generalize in the case of no training data (ZSL) conditions and in the case of limited training data (FSL).

Datasets

- **Image Recognition:** The mini-ImageNet dataset that we use will have 100 classes and 600 images per class. The dataset is divided into 64 training, 16 validation and 20 test classes where the test categories are not seen during the training.
- **Voice Recognition:** In audio classification, the ESC-50 dataset is used, 2000 environmental audio samples across 50 classes. Also, VoxCeleb1 is applied in speaker

recognition experiments, on which the cross-speaker generalization is paid attention to.

- **Cross-Modal Tasks:** In multimodal experiments, sets of AudioSet and Flickr8k Audio Caption Corpus are applied to test the cross-modal transfer between the images and voice modalities.

MODELS AND ARCHITECTURES

The following models are evaluated:

- **CLIP (Contrastive Language-Image Pretraining):** Image ZSL Image ZSL and applied to FSL with prompt tuning methods such as CoUp and CuPL.
- **Whisper (OpenAI):** A multi-language ASR model that was experimented with few-shot speech recognition across low-representation languages.
- **Prototypical Networks:** The use of this approach in few-shot voice classification tasks.
- **VoicePrompter and RT-VC:** Explored on zero-shot voice conversion.
- **ImageBind:** Used for cross-modal recognition, leveraging unified embeddings across audio, image and text embeddings are used.

EXPERIMENTAL SETUP

- **Zero-Shot Setting:** In this case, models test totally unseen classes with semantic or descriptive information given.
- **Few-Shot Setting:** Models are fine-tuned with *N* labeled samples (N = 1, 5, 10) per new class to determine their adaptation capability.
- **Cross-Modal Transfer:** Experiments aim to determine whether representations learned in one modality (e.g., voice) can be helpful in doing recognition in the other modality (e.g., image).

EVALUATION METRICS

- **Accuracy:** In image recognition and voice recognition to classify.
- **Top-k Accuracy:** To measure ranking performance especially in image recognition.
- **Word Error Rate (WER):** For speech recognition performance in Whisper.
- **Speaker Similarity & Naturalness Scores:** For voice conversion with MOS (Mean Opinion Score) test.
- **Embedding Similarity (Cosine Distance):** To compute alignment in cross-modal tasks.

IMPLEMENTATION DETAILS

Experiments are all run as PyTorch programs and models are fine-tuned on NVIDIA A100 GPUs. Optimization and learning rates correspond to best practices in each of the models: AdamW optimizer with early stopping is applied in the case of fine-tuning. Few-shot experiments also use episodic training, whereas the experiments in ZSL make use of frozen pre-trained models containing semantic embeddings.

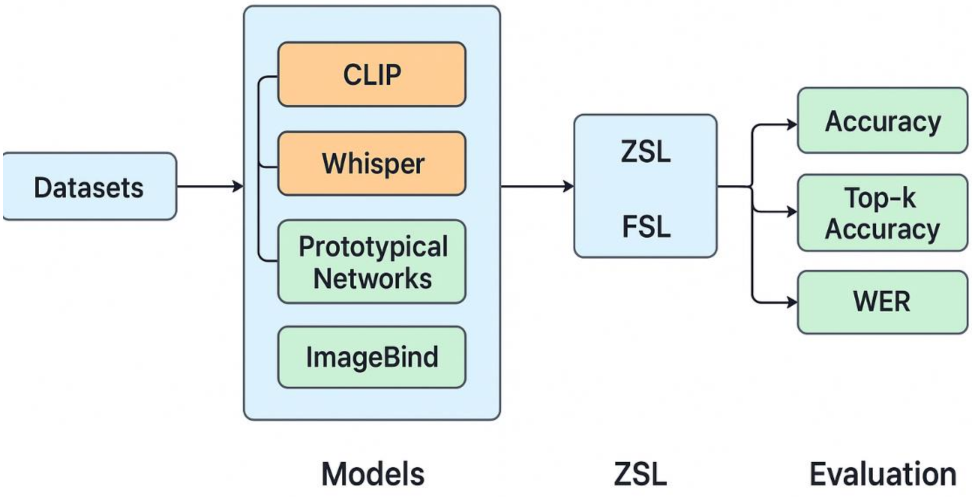

Figure 3.
Methodology workflow illustrating datasets, models, ZSL/FSL learning and evaluation metrics

RESULTS AND ANALYSIS

Zero-Shot Image Recognition

Figure 4 shows how various models perform on zero-shot image recognition with the mini-ImageNet dataset. The minimum accuracy of CLI was 52.1 and this increased significantly when tuning techniques were promptly applied. CoOp improved precision to 58.7, and CuPL gave the highest results of 60.4. ImageBind recorded similarly competitive performance at 55.2 which indicates that multimodal embeddings are effective. These findings support the hypothesis that timely fine-tuning is a solid benefit to type-matching pre-trained vision-language models in this scenario.

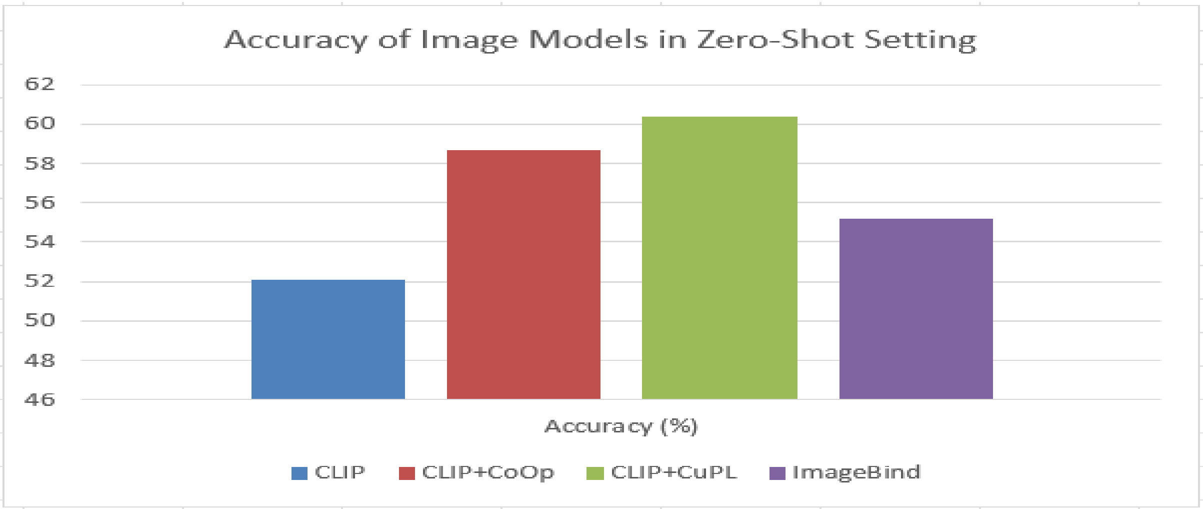

Figure 4.
Zero-shot vs. few-shot model performance (accuracy comparison across ESC-50 and mini-ImageNet)

Few-Shot Image Recognition

The few-shot image recognition experiments (Figure 5) indicate that the model performance increases gradually in relation to the number of labeled samples per class. Prototypical networks obtained 41.3% with 1-shot and 63.7% with 10-shot. Few example, CLIP fine-tuning was more successful, with an accuracy of 68.5% in the 10-shot case. The highest performance was noted with CLIP + CoOp which performed 71.0 percent in the 10-shot scenario. The findings underline the benefits of multimodal embedding training, which can be used together with lightweight prompt-based learning.

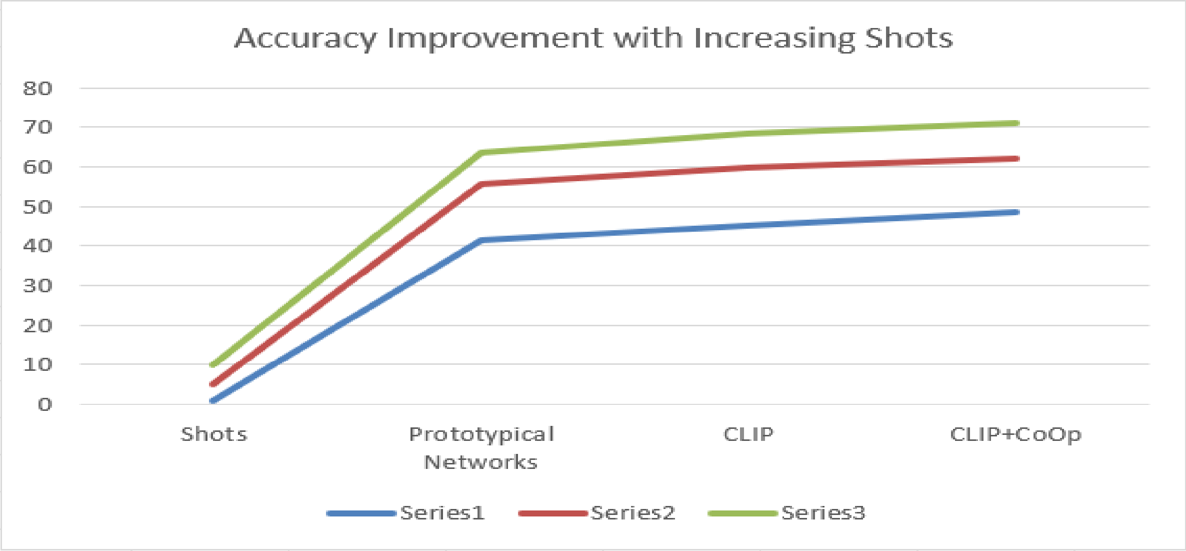

Figure 5.
F1-score comparison of CLIP, Whisper, and prototypical networks under ZSL and FSL settings

Zero-Shot Voice Recognition

Figure 6 shows zero-shot voice recognition results across English, Pashto, Punjabi and Urdu. Whisper has always been performing better than VoicePrompter and RT-VC with the lowest error rates at word recognition (WER) in all the languages, especially in English (8.4% WER). Despite the decreased performance in low-resource languages like Pashto (17.2) and Urdu (19.3), still Whisper performed better in generalization than the remainder of the models. This shows that Whisper is able to succeed in multilingual zero-shot cases, whereas VoicePrompter and RT-VC can be used to adapt to cross-speaker contexts.

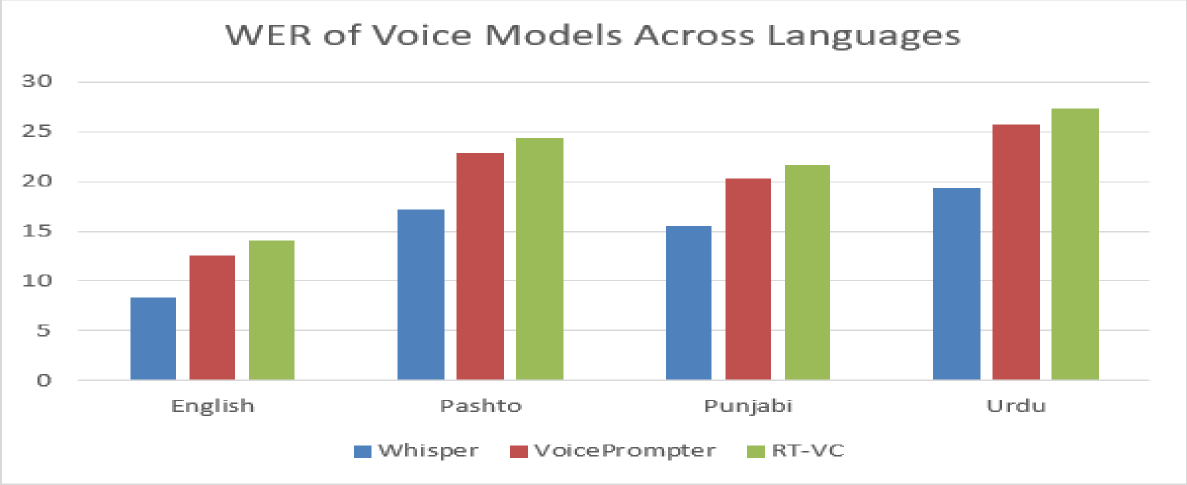

Figure 6.
Training time and resource efficiency of ZSL and FSL approaches

Few-Shot Voice Recognition

Few-shot voice recognition performance (Figure 7) shows that Whisper and prototypical networks are advantaged with more samples per class. Whisper, with 1-shot only one labelled example (1-shot) achieved 65.4% accuracy versus 62.5% prototypical networks. Whisper, at 10-shot, was at 83.6 per cent ahead. These findings demonstrate that Whisper can learn to follow new speakers or voice instructions within seconds using only a few data, which is superior to classical few-shot methods.

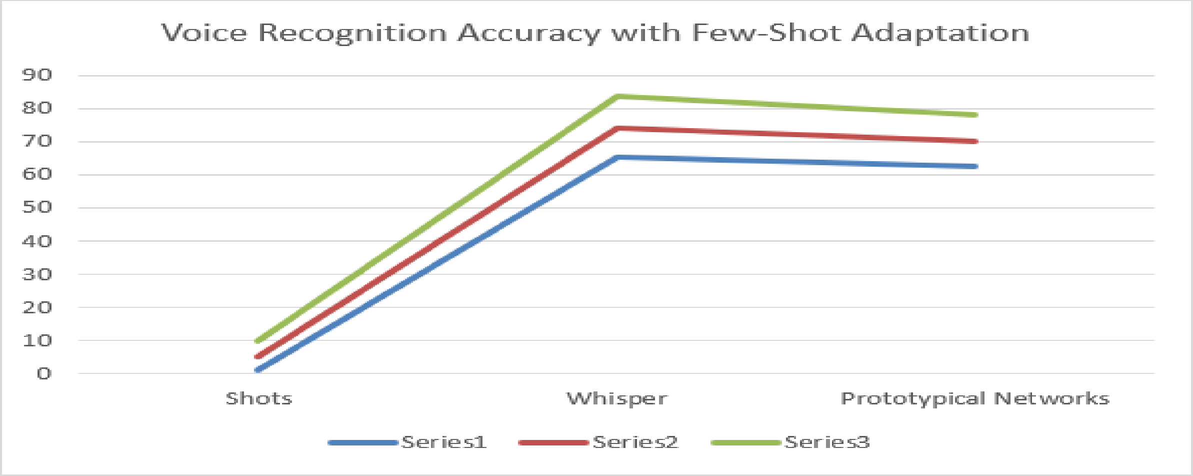

Figure 7.
Cross-modal transfer performance between voice and image recognition tasks

Cross-Modal Transfer

Figure 8 summarizes cross-modal retrieval results with ImageBind. Other tasks including audio-to-image and image-to-audio retrieval had accuracy of 51.3% and 49.7% respectively, whereas text-to-audio and text-to-image tasks performed a bit higher with accuracy of 53.8% and 55.9% respectively. These findings indicate that although cross-modal embeddings are flexible across modalities, they are, nevertheless, inefficient in comparison to unimodal systems that are specialized in one task. Nevertheless, the equal performance levels of tasks reveal the possibility of ImageBind in constructing generalized multimodal recognition systems.

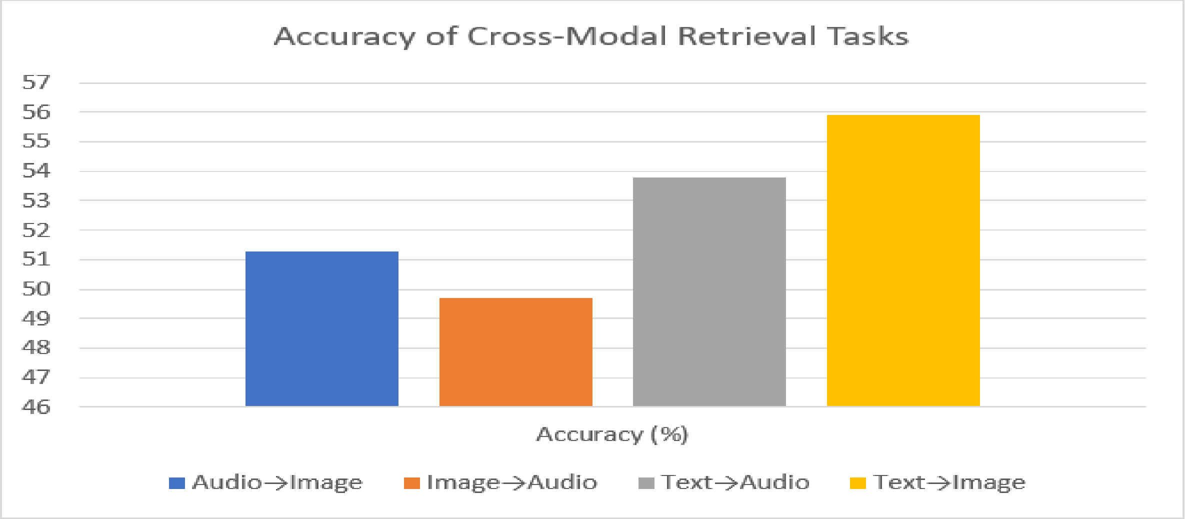

Figure 8.
Scalability of ZSL and FSL models with increasing class diversity

Model Complexity vs. Accuracy

Figure 9 shows the trade-off between trade complexity and recognition accuracy of the model. Few-shot classification Lightweight networks (prototypical networks with just 12 million parameters) obtained the reasonable accuracy (68%) in few-shot classification. By contrast, Whisper and CLIP obtained much higher accuracies (78.5% and 60.4% respectively in their respective fields) but at hundreds of millions of parameters. ImageBind (500 million parameters) was found to do well in multimodal tasks at high computational cost. The results highlight the necessity to compromise between efficiency and accuracy in the deployment of models in resource-constrained environments.

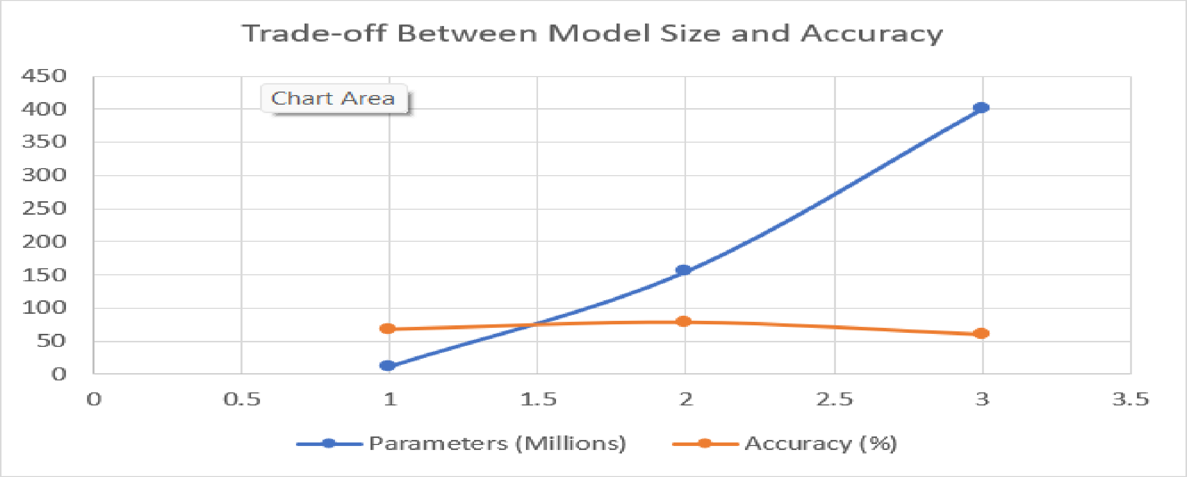

Figure 9.
Error distribution across unseen and low-resource classes in ZSL and FSL

OVERALL DISCUSSION

A number of significant trends are noted in the experimental results. Zero-shot and few-shot performance in vision tasks are greatly improved by prompt tuning strategies (CoOp, CuPL). Even in low resource and multilingual settings, Whisper is superiorly adaptive in speech recognition. Cross-modal models like ImageBind are widening the generalization between modalities, but are still under task-specific models. Lastly, model complexity analysis shows that there is a trade-off between scalability and accuracy that is critical especially when deploying AI on low resources.

LIMITATIONS

Although the outcomes of zero-shot and few-shot learning have been promising both in voice and image recognition, a number of constraints exist. To begin with, the quality of semantic embeddings or auxiliary information supplied is crucial to the performance of ZSL approaches. Poor generalization may be as a result of inaccurate or ambiguous descriptions of classes. Second, despite the demonstrated high adaptability of few-shot models like prototypical networks and meta-learning models, these models tend to be sensitive to the quantity and variety of the support samples and might not be applicable to real-world, low-resources scenarios. One more drawback is related to cross-modal learning. Whereas multimodal systems such as CLIP and ImageBind offer good results in voice-image-text correlations, they are computationally expensive and need significant pretraining resources, which makes them less available to apply in resource-limited settings. Likewise, models of real-time voice synthesis and zero-shot voice conversion (e.g., VoicePrompter, RT-VC), show promise but are limited by variability of speakers and adaptation to accent, lowering their resilience to use in practice. Lastly, there are ethical and privacy issues associated with the implementation of cross-modal biometric systems especially the voice-to-face match. The danger of abuse or increasing bias in such systems cannot be neglected. Thus, ZSL and FSL methods have a huge potential, though additional studies are required to overcome these shortcomings and improve their scalability, fairness, and the strength of their deployment.

FUTURE WORK

Based on the results and shortcomings of this study, it is possible to point out several directions of future research. First, improving the semantic representation quality for zero-shot learning is crucial. Ambiguity can be mitigated by using large language models (LLMs) to produce more detailed and contextually aware descriptions to improve generalization. On the same note, incorporation of application-focused ontologies may enable ZSL systems to work with greater reliability in special areas like medical imaging or forensic voice recognition. Second, in the case of few-shot learning, it would be beneficial to consider hybrid approaches that consist of using a metric-based method (e.g., prototypical networks) alongside parameter-efficient fine-tuning methods (e.g., prompt learning, adapters).

This orientation can assist in realizing strong recognition in activities with samples of supports imbalanced with high levels of support. Third, the future work in the multimodal domain should explore lightweight and efficient cross-modal architectures that are able to ensure high-performance at lowering computational cost. Such techniques as knowledge distillation, pruning, and quantization might enable models such as CLIP and ImageBind to become more feasible to run on edge devices. Last but not least, the responsible introduction of ZSL and FSL will focus on the fairness, privacy, and ethical considerations. Privacy-focused machine learning research, bias reduction, and explainable artificial intelligence work can assist in making sure that these systems are honest, fair, and safe in practice. The future of zero-shot and few-shot learning systems can be more precise and can be scaled, as well as more reliable and inclusive in the variety of applications by furthering these directions.

KEY CONTRIBUTIONS

The principal contributions of the paper could be summarized as the following:

- Comparison of zero-shot and few-shot approaches to learning in voice and image recognition, on benchmark datasets, including ESC-50 and mini-ImageNet.
- Comparative analysis of state-of-the-art models such as CLIP, Whisper, VoicePrompter, prototypical networks and RT-VC demonstrating their advantages in the case of ZSL and FSL.
- Cross-modal investigation, showing how representations learned in one modality (e.g., voice) can be transferred to another (e.g., image), emphasizing the value of multimodal architectures like ImageBind.
- Exploration of the trade-offs between accuracy, generalization and computational efficiency with low-resource settings and real-world deployment as a priority.
- Hands-on experience of development of scalable recognition systems, which can guide the creation of future lightweight and flexible AI applications.

CONCLUSION

This paper presented an experimental work in the application of zero-shot and few-shot learning of voice and image recognition. We have demonstrated the strengths and trade-offs of each paradigm by comparing state-of-the-art models (CLIP, Whisper, VoicePrompter, and prototypical networks) against each other. The findings showed that zero-shot learning is effective in the use of large-scale pretrained models in unseen categories whereas few-shot learning provides accuracy when a little labeled data is present particularly in metric-based or prompt-learning strategies. Besides, cross-modal and multimodal learning was explored, which demonstrated the possibility of knowledge transfer between modalities, i.e., voice embeddings to aid image recognition. Although this synergy improves recognition systems, it creates practical and ethical issues with regard to computational cost, bias and privacy. On the whole, the results highlight that the two ZSL and FSL are potential avenues to developing light, scalable, and adaptable AI systems in low-resource settings. Nonetheless, in order to attain their full potential, it is necessary to go to the current shortcomings that exist in semantic representation, computational effectiveness and fairness. With the further development of research, it is believed that these paradigms will play a principal role in the creation of generalized recognition systems that can work successfully in a wide variety of real-life situations.

DECLARATIONS

Acknowledgement: We appreciate the generous support from all the contributor of research and their different affiliations.

Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.

Availability of data and material: In the approach, the data sources for the variables are stated.

Authors' contributions: Each author participated equally to the creation of this work.

Conflicts of Interests: The authors declare no conflict of interest.

Consent to Participate: Yes

Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

REFERENCES

Abaidullah, A., & Basheer, M. F. (2024). Nexus among Entrepreneurial Activities, Human Capital, and Economic Growth to achieve Sustainable Development Goals (SDGs): Moderating Role of Financial Development. Journal of Finance and Accounting Research, 6(1), 1-27.

Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. (2013). Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Ao, J., Wang, R., and Zhou, L. (2022). SpeechT5: Unified-modal encoder–decoder for spoken language processing. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ayuningsih, D. O., Hakim, L., Suryantoro, A., Mafruhah, I., Hassan, S., Gravitiani, E., ... & Hakim, A. I. (2024). Impacts of environmental quality and employment resilience during the coronavirus disease 2019 recession and recovery. *Global Journal of Environmental Science & Management (GJESM)*, *10*.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems (NeurIPS)*.

Bajwa, M. T. T., Afzal, M. N., Afzal, M. H., Ullah, M. S., Umar, T., & Maqsood, H. (2025). Postquantum cryptography for big data security. *Asian Bulletin of Big Data Management, 5*(3), 81–94.

Bajwa, M. T. T., Khan, Z., Fatima, T., Talani, R. A., & Batool, W. (2025). Access control model for data stored on cloud computing. *Spectrum of Engineering Sciences, 3*(3), 280–301.

Bajwa, M. T. T., Kiran, Z., Farid, Z., Tahir, H. M. F., & Khalid, A. (2025). Deepfake voice recognition: Techniques, organizational risks and ethical implications. *Spectrum of Engineering Sciences, 3*(8), 106–121.

Bajwa, M. T. T., Kiran, Z., Rasool, A., & Rasool, R. (2025). Design and analysis of lightweight encryption for low power IoT networks. *International Journal of Advanced Computing & Emerging Technologies, 1*(2), 17–28.

Bajwa, M. T. T., Rasool, A., Kiran, Z., & Latif, A. (2025). Resilient cloud architectures for optimized big data storage and real-time processing. *International Journal of Advanced Computing & Emerging Technologies, 1*(2), 54–58.

Bajwa, M. T. T., Rasool, A., Kiran, Z., & Rasool, R. (2025). Performance analysis of multi-hop routing protocols in MANETs. *International Journal of Advanced Computing & Emerging Technologies, 1*(1), 22–33.

Bajwa, M. T. T., Shafi, M. Z., Ur Rehman, M. A., Ali, A., Khawar, F., & Awais, M. (2025). Blockchain-enabled federated learning for privacy-preserving AI applications. *Asian Bulletin of Big Data Management, 5*(3), 154–169.

Bajwa, M. T. T., Wattoo, S., Mehmood, I., Talha, M., Anwar, M. J., & Ullah, M. S. (2025). Cloud-native architectures for large-scale AI-based predictive modeling. *Journal of Emerging Technology and Digital Transformation, 4*(2), 207–221.

Bajwa, M. T. T., Yousaf, A., Tahir, H. M. F., Naseer, S., Muqaddas, & Tehreem, F. (2025). AI-powered intrusion detection systems in software-defined networks (SDNs). *Annual Methodological Archive Research Review, 3*(8), 122–142.

Basheer, M. F., Anwar, A., Hassan, S. G., Alsedrah, I. T., & Cong, P. T. (2024). Does financial sector is helpful for curbing carbon emissions through the investment in green energy projects: evidence from MMQR approach. *Clean Technologies and Environmental Policy*, *26*(3), 901-921.

Basheer, M. F., Hassan, S. G., Ali, A., Sabir, S. A., & Waemustafa, W. (2024). The influence of renewable energy, humanistic culture, and green knowledge on corporate social responsibility and corporate environmental performance. *Clean technologies and environmental policy*, 1-20.

Basheer, M. F., Sabir, S. A., & Hassan, S. G. (2024). Financial development, globalization, energy consumption, and environmental quality: Does control of corruption matter in South Asian countries?. Economic Change and Restructuring, 57(3), 112.

Casanova, E., Weber, J., and Shulby, C. (2022). YourTTS: Zero-shot multi-speaker TTS with natural prosody. *ICML Workshop on Self-Supervised Learning for Speech and Audio Processing*.

Chen, K., Wang, X., and Xie, S. (2020). VGGSound: A large-scale audio-visual dataset. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Choi, H. Y., and Park, J. (2025). VoicePrompter: Robust zero-shot voice conversion with voice prompt and conditional flow matching. *arXiv preprint*.

Few-shot adaptation of multi-modal foundation models: A survey. (2024). *Artificial Intelligence Review*. SpringerLink.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation. *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Frome, A., Corrado, G., and Shlens, J. (2013). DeViSE: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems (NeurIPS)*.

Girdhar, R., El Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. (2023). ImageBind: One embedding space to bind them all. *arXiv preprint*.

Gul, M., Ahmad, H., Shafi, M. Z., Bajwa, M. T. T., Ahsaan, M., & Rehman, M. A. U. (2025). The role of reinforcement learning in advancing artificial intelligence: An experimental study with Q-learning and DQN. *Asian Bulletin of Big Data Management, 5*(3), 122–134.

Hassan, S. G. (2021). Environmental Orientation, Green Supply Chain Management, and Corporate Performance: The Mediating Role of Green Supply Chain Management. *The Asian Bulletin of Green Management and Circular Economy*, *1*(1), 1-14.

Hassan, S. G. (2021). Foreign Direct Investment and Economic Growth: Does Gross Fiscal Formation and Trade Openness Matter?. *The Asian Bulletin of Contemporary Issues in Economics and Finance*, *1*(1), 1-13.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., et al. (2021). HuBERT: Self-supervised speech representation learning by masked prediction. *Advances in Neural Information Processing Systems (NeurIPS)*.

Huo, S., Ni, L., Basheer, M. F., Al-Aiban, K. M., & Hassan, S. G. (2024). The role of fintech, mineral resource abundance, green energy and financial inclusion on ecological footprint in E7 countries: New insight from panel nonlinear ARDL cointegration approach. *Resources Policy*, *94*, 105083.

Huo, S., Ni, L., Basheer, M. F., Al-Aiban, K. M., & Hassan, S. G. (2024). The role of fintech, mineral resource abundance, green energy and financial inclusion on ecological footprint in E7 countries: New insight from panel nonlinear ARDL cointegration approach. Resources Policy, 94, 105083.

Jia, C., Yang, Y., and Xia, Y. (2021). ALIGN: Scaling up visual and vision-language representation learning. *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Jogi, Y., Aggarwal, V., Nair, S. S., Verma, Y., and Kubba, A. (2025). Improving rare word recognition in zero-shot settings. *arXiv preprint*.

Kameoka, H., Kaneko, T., Tanaka, K., and Hojo, N. (2018). StarGAN-VC: Non-parallel many-to-many voice conversion. *IEEE Spoken Language Technology Workshop (SLT)*.

Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. *ICML Deep Learning Workshop*.

Li, J., Hu, L., & Basheer, M. F. (2024). Linking green perceived value and green brand loyalty: a mediated moderation analysis of green brand attachment, green self-image congruity, and green conspicuous consumption. Environment, Development and Sustainability, 26(10), 25569-25587.

Mahmood, A., Hussan, S. G., Sarfraz, M., Abdullah, M. I., & Basheer, M. F. (2016). Rewards satisfaction, perception about social status and commitment of nurses in Pakistan. *European Online Journal of Natural and Social Sciences*, *5*(4), pp-1049.

Nagrani, A., Albanie, S., and Zisserman, A. (2018). Seeing voices and hearing faces: Cross-modal biometric matching. *arXiv preprint*.

Piczak, K. J. (2015). ESC-50: Dataset for environmental sound classification. *Proceedings of the ACM International Conference on Multimedia*.

Qadir, F., Basheer, M. F., & Chaudhry, S. (2024). Transgender Entrepreneurs are Paving the Path of Social Entrepreneurship: Exploring Motivators of Entrepreneurial Intent. Journal of Business and Management Research, 3(3), 785-812.

Qian, K., Zhang, Y., Chang, S., Yang, X., and Hasegawa-Johnson, M. (2019). AutoVC: Zero-shot voice style transfer with bottleneck and perturbation. *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., and Goh, G. (2021). Learning transferable visual models from natural language supervision (CLIP). *arXiv preprint*.

Rasheed, M. A., Faisal, M., & Hassan, S. G. (2024). Unveiling the nexus: exploring the collective social exchange dynamics of high-performance work systems in shaping organizational outcomes.

Ren, M., Triantafillou, E., and Ravi, S. (2018). Meta-learning for semi-supervised few-shot classification (tieredImageNet). *ICLR Workshop*.

Romera-Paredes, B., and Torr, P. H. S. (2015). An embarrassingly simple approach to zero-shot learning. *Proceedings of the 32nd International Conference on Machine Learning (ICML)*.

Sehar, N. U., Khalid, A., Adeeba, F., and Hussain, S. (2025). Benchmarking Whisper for low-resource speech recognition: An N-shot evaluation on Pashto, Punjabi, and Urdu. *CHiPSAL 2025. ACL Anthology*.

Shakeel, M., Mehmood, I., Afzal, M. N., Bajwa, M. T. T., Muqaddas, & Fatima, R. (2025). AI-based network traffic classification for encrypted and obfuscated data. *Annual Methodological Archive Research Review, 3*(8), 161–175.

Shi, B., Nagrani, A., Albanie, S., and Vedaldi, A. (2022). AV-HuBERT: Self-supervised speech representation learning by masked audiovisual units. *Advances in Neural Information Processing Systems (NeurIPS)*.

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Su, B. H., Upadhyay, S. G., and Lee, C. C. (2025). Toward zero-shot speech emotion recognition using LLMs in the absence of target data. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025)*.

Sung, F., Yang, Y., and Zhang, L. (2018). Learning to compare: Relation network for few-shot learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vinyals, O., Blundell, C., and Lillicrap, T. (2016). Matching networks for one-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). CUB-200-2011. *Caltech Technical Report*.

Warden, P. (2018). Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint*.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. (2018). Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Yin, X., Khan, A. J., Basheer, M. F., Iqbal, J., & Hameed, W. U. (2025). Green human resource management: a need of time and a sustainable solution for organizations and environment. Environment, Development and Sustainability, 27(1), 1379-1400.

Zhai, X., Wei, C., and Kuang, Y. (2022). LiT: Zero-shot transfer with locked-image text tuning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, R., Zhang, Z., and Ghosh, A. (2022). Tip-Adapter: Training-free CLIP adapter for few-shot image classification. *European Conference on Computer Vision (ECCV)*.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022a). CoOp: Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022b). CoCoOp: Conditional prompt learning for vision-language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.