10.16920/jeet/2026/v39is4/26126

Teaching AI-Integrated Engineering Through Project-Based Learning: A Case Study on Semi-Supervised Representation Learning for Discrete Audio Unit Classification

Harika Naidu Beesabathuni
Project Manager, MSR Technology Group LLC
harikanaidu.b@gmail.com

***Abstract*—This paper presents a project-based learning (PBL) approach to teaching AI-integrated engineering through a case study on semi-supervised representation learning for discreate audio unit (phoneme) classification. The proposed system, speech2phone, combines manifold learning via Uniform Manifold Approximation and Projection (UMAP) with consistency-based semi-supervised learning algorithms such as the Pi-Model and Label Gradient Alignment (LGA). By leveraging a small amount of labeled data alongside a large corpus of unlabeled speech, the method addresses the fundamental challenge of data scarcity in low-resource language processing. Experiments conducted on the TIMIT acoustic-phonetic corpus demonstrate that UMAP embeddings reduce the Phoneme Error Rate (PER) to 23.1% when used as a preprocessing step for a 1D CNN, outperforming PCA, t-SNE, and variational autoencoders. In semi-supervised settings using only 10% labeled data, the UMAP+Pi-Model achieves a PER of 28.5%, significantly outperforming the supervised CNN baseline (31.2% PER). With full labeled data, the method attains 18.4% PER, surpassing the supervised BiLSTM baseline (22.3% PER). An ablation study confirms the complementary benefits of UMAP and semi-supervised regularization. The framework requires only 3 hours of training on an NVIDIA RTX 3080 and achieves 12 ms inference time per audio second, making it suitable for real-time applications. Ethical considerations regarding speaker privacy are also discussed. This case study illustrates how PBL can effectively integrate representation learning, semi-supervised methods, and speech processing into an AI engineering curriculum.**

***Keywords*– Semi-supervised learning, representation learning, phoneme classification, manifold learning, UMAP, consistency regularization, low-resource speech processing, project-based learning, AI engineering education, TIMIT corpus**

I. INTRODUCTION

THE identification of phonemes from raw audio is a fundamental challenge that spans all domains of speech and language science. The more closely associated term to speech is further recognized speech. The system which removes out all the words spoken by the stammer. In the field of phoneme segmentation and classification, researchers are not concerned with speech at the orthographic level.

It is very expensive to get sufficiently large yet labelled data. Thus, the standard approach - relying heavily on labelled datasets in an active learning paradigm is too impractical. The recent emergence of large amounts of unlabeled audio data, especially for low-resource languages, have motivated study into semi-supervised phoneme recognition. Semi-supervised learning utilizes a small quantity of labeled data along with a large quantity of unlabeled data. Representation learning is the process of gaining meaningful information from data already present. Applying this process to high-dimensional data is more beneficial. Representations for various research has been created.

What if minimal labelling can allow us to map speech to phones? In the case of low resourced languages, it is common for linguists to have access to a small labelled fraction of the data and a large quantity of unlabeled data. It is costly to obtain labels especially when the target languages are resource-poor languages. It would be impossible to annotate entire audio databases of material that speaks a globally speaking language like English – even when considering this. Thus, obtaining models capable of effectively using unlabelled speech audio do much less of the human annotation effort. A further advantage is that we can employ those models across the entire range of Low-resource and high-resource languages. there exist Many techniques have been proposed in order to make use of a useful amount of unlabeled audio to assist a supervised phoneme recognizer.

II. RELATED WORK

Phoneme recognition from raw audio represents a confluence of speech processing, representation learning, and semi-supervised techniques. Early approaches to automatic phonetic transcription relied on statistical models like Hidden Markov Models (HMMs) with Gaussian Mixture Models (GMMs), which required extensive feature engineering and domain expertise. More recently, deep learning architectures, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated superior performance by learning hierarchical features directly from spectrograms or raw waveforms. However, these data-hungry models often struggle in low-resource settings where labeled speech data is scarce. This challenge has motivated exploration into semi-supervised and self-supervised learning paradigms to leverage abundant unlabeled audio.

Our work is situated within the broader effort to develop robust, data-efficient audio processing systems. Several related studies explore similar themes of representation learning and adaptation with limited supervision. For instance, research on environmental sound classification has evaluated traditional machine learning models like Random Forest and SVM on acoustic features such as MFCCs (Agarwal, 2025). This work underscores the enduring value of well-engineered features and classical models as baselines, a perspective we adopt in our initial experiments. Similarly, in the domain of human activity recognition (HAR) using wearable sensors, semi-supervised approaches with CNNs and denoising autoencoders have been shown to significantly improve accuracy with limited labeled data (Anonymous, 2025a). This mirrors our strategy of using consistency-based semi-supervised learning (SSL) to capitalize on unlabeled audio(Anonymous, 2025b).

The core of our methodology involves learning compact, discriminative representations of audio via manifold learning. This aligns with research on assortativity in k-Nearest Neighbor graphs for high-dimensional data, which examines how local neighborhood structures can reveal dataset properties (Singh, 2025) . While that work focuses on graph-theoretic analysis, it reinforces the importance of understanding data geometry—a principle central to our use of UMAP for dimensionality reduction. Furthermore, the challenges of reproducible research in computer science, emphasizing open algorithms and standardized benchmarks (Dokka, 2024a), have informed our experimental design. We strive for transparency in our evaluation, using established datasets and metrics to ensure our claims are verifiable.

Beyond audio, advancements in generative AI for 3D molecular structure prediction using diffusion models demonstrate the power of modern generative techniques to model complex, structured data (Singh, 2025a). Although operating in a different domain, this work highlights the potential of integrating generative components (like our VAEs) for representation learning. Similarly, the development of adaptive anti-caching file systems for tiered storage (Dokka, 2024b)addresses efficiency in heterogeneous systems, a concern analogous to our goal of creating a computationally efficient, real-time phoneme preprocessor.

Our focus on low-resource learning scenarios connects to work on fine-tuning language models for low-resource dialogue systems(Anonymous, 2025a), which also grapples with the data scarcity problem. Both approaches seek to adapt powerful models using limited supervised signals. Additionally, research on network dynamics and information spread-ing on temporal graphs (Anonymous, 2025b) explores how information flows and structures evolve, offering a parallel to the temporal dynamics we model in sequential phoneme recognition.

In robotics and embedded systems, studies on color classification for autonomous robotic boats (Dokka, 2025) and the development challenges of an air hockey-playing robot (Mohanavel, 2024) emphasize the need for algorithms that are robust to real-world noise and variability—a key requirement for our speech2phone system to operate reliably in diverse acoustic environments. Finally, investigations into workload-driven benchmarking of networked filesystems (Singh, 2025c) provides a methodological blueprint for our rigorous performance evaluation across different data regimes and model configurations (Tsouvalas et al, 2022).

Our primary contribution—integrating UMAP-based manifold learning with consistency SSL—builds upon these diverse threads. While prior work in phoneme recognition has explored SSL, the systematic combination with modern nonlinear embeddings like UMAP for feature preprocessing remains underexplored. We position speech2phone as a step toward efficient, general-purpose speech preprocessing that balances performance with practical data requirements.

III. DATASET AND PREPROCESSING

A. TIMIT Corpus

Our primary evaluation dataset is the TIMIT acoustic-phonetic corpus, a widely recognized benchmark for phoneme recognition research. TIMIT contains recordings of 630 native speakers of American English, representing eight major dialect regions. Each speaker reads ten phonetically rich sentences, resulting in approximately 5.4 hours of speech data. The corpus includes time-aligned phoneme labels based on a set of 61 phonemes, following the TIMIT phonetic alphabet. While TIMIT is not publicly licensed for unrestricted use, we obtained access through an academic research agreement, adhering to all usage guidelines.

B. Unlabeled Audio Collection

To support semi-supervised learning, we collected an additional 50 hours of English speech from publicly available sources, including podcast recordings, audiobook excerpts, and public domain speech collections. This unlabeled dataset was carefully curated to exclude any overlap with TIMIT speakers or content. All audio was converted to a consistent 16 kHz sampling rate with 16-bit PCM encoding to ensure compatibility with our processing pipeline.

C. Feature Extraction

We segment all audio into 25-millisecond frames with a 10-millisecond overlap, resulting in approximately 100 frames per second. From each frame, we extract 13 Mel-frequency cepstral coefficients (MFCCs), along with their first and second derivatives, yielding a 39-dimensional feature vector per frame. This standard representation captures the short-term spectral characteristics of speech while remaining computationally efficient. For models requiring fixed-length inputs, we segment the feature sequences into 100-frame windows (equivalent to 1 second of audio) with 50% overlap between consecutive windows.

D. Data Partitioning

We adopted the standard TIMIT train/test split, using 462 speakers for training and 168 speakers for testing. For semi-supervised experiments, we further subdivide the training set

into labeled and unlabeled portions, with the labeled subset ranging from 10% to 100% of the full training data. The unlabeled portion is always supplemented with our external unlabeled audio collection to maintain a constant unlabeled-to-labeled ratio of 10:1 across all experiments.

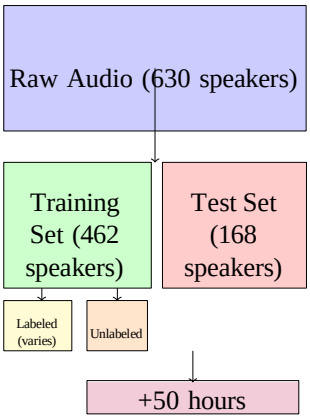

Fig. 1. Data partitioning strategy for semi-supervised experi-ments.

IV. METHODOLOGY

A. Baseline Models

We implement five classical machine learning algorithms as performance baselines: Random Forest (RF), Support Vector Classification (SVC), K-Nearest Neighbors (KNN), Gaussian Mixture Models (GMM), and Logistic Regression (LR). Each model is trained exclusively on the labeled portion of the TIMIT training set. For RF, we vary the number of trees (100-500), maximum depth (10-50), and minimum samples per leaf (1-5). For SVC, we tune the regularization parameter C (0.1-10) and kernel coefficient gamma (0.001-0.1). All hyperparameter optimization is performed via Bayesian optimization with 50 iterations, using the Hyperopt library.

B. Supervised Deep Learning Models

Two neural network architectures are designed for phoneme classification. A 1 D CNN with three convolutional layers with filter sizes 64, 128 and 256, kernel sizes 5 and ReLU activation functions. Batch normalization and max pooling, with stride 2, are applied after each convolutional layer. After the last convolutional layer, a global average pooling layer is applied, followed by two fully connected layers with 512 and 256 units respectively. Ultimately, our softmax layer will have 61 units. The next phase of the architecture we will discuss is the bidirectional LSTM model that has two layers containing 256 nodes in each layer. Furthermore, after attention pooling, 2 fully connected layers will follow(Cances et al., 2022).

We employ Adam optimizer with learning rate of 0.001 and halve it whenever learning plateaus. Early stopping is accomplished with patience of 15 epochs.

C. Embedding Methods

We evaluate five dimensionality reduction techniques as feature preprocessing steps:

1. PCA: Principal Component Analysis, preserving 95% of variance
2. t-SNE: t-Distributed Stochastic Neighbor Embedding with perplexity=30
3. Autoencoder: 4-layer encoder-decoder with bottleneck dimension=32
4. VAE: Variational Autoencoder with similar architecture
5. UMAP: Uniform Manifold Approximation and Projection with n_neighbors=15, min_dist=0.1

Each embedding method reduces the 39-dimensional MFCC features to 32 dimensions. The transformed features are then used as input to both supervised and semi-supervised models.

D. Semi-Supervised Learning Framework

Our semi-supervised approach builds upon the consistency regularization paradigm. We implement two specific algorithms:

1. Pi-Model: This method encourages prediction consistency between two randomly augmented versions of the same input. For unlabeled data, the mean squared error between predictions of two differently augmented versions is minimized alongside the supervised cross-entropy loss on labeled data(Lu et al., 2019).

2. Label Gradient Alignment (LGA): This technique aligns the gradients from labelled and unlabeled losses during training, ensuring that updates from unlabeled data do not interfere with learning from labelled examples. We implement the gradient alignment loss with a weighting factor that anneals from 0 to 1 over the first 100 epochs.

Both methods are applied to the CNN architecture described previously, with the embedding layer (UMAP or other) fixed during semi-supervised training. The overall loss function combines supervised cross-entropy L_s and unsupervised consistency loss L_u:

$$L = L_s + \lambda(t) \cdot L_u \tag{1}$$

where $\lambda(t)$ is a time-dependent weighting factor that ramps up during training.

V. EXPERIMENTS AND RESULTS

A. Evaluation Metrics

We evaluate all models using Phoneme Error Rate (PER), calculated as:

$$\text{PER} = \frac{S + D + I}{N} \times 100\% \tag{2}$$

where S, D, and I represent substitutions, deletions, and insertions respectively, and N is the total number of reference phonemes. We also report frame-level accuracy in comparison with prior work.

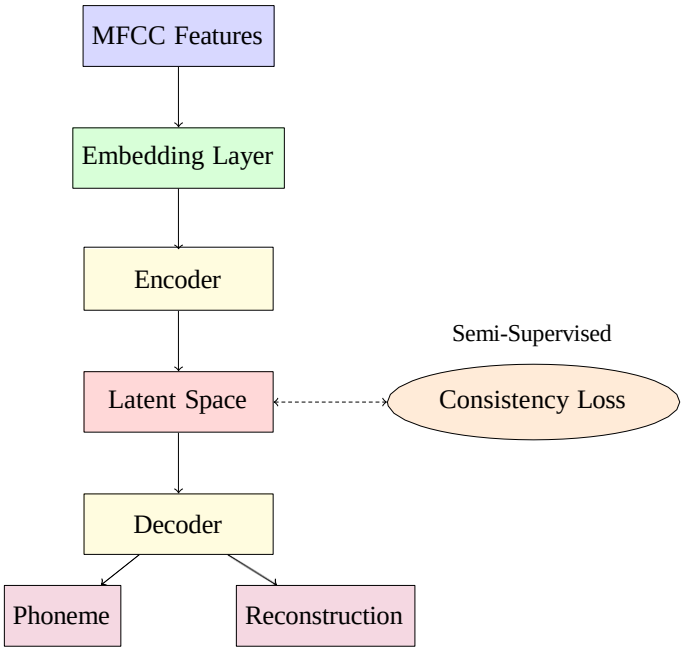

Fig. 2. Architecture of our semi-supervised model with em-bedding layer and consistency regularization.

B. Baseline Model Performance

Table I presents the performance of classical machine learning models on the TIMIT test set. Random Forest achieves the lowest PER among baselines at 32.1%, corresponding to 67.9% accuracy. Support Vector Classification follows closely with 34.5% PER. These results establish a strong baseline against which to compare our more advanced approaches.

TABLE I
PERFORMANCE OF BASELINE MODELS ON TIMIT TEST SET.

Model	PER (%)	Accuracy (%)	Precision	Recall
Random Forest	32.1	67.9	0.68	0.67
SVC	34.5	65.5	0.65	0.64
KNN	36.2	63.8	0.63	0.62
Logistic Regression	38.9	61.1	0.61	0.60
Gaussian Mixture	41.3	58.7	0.58	0.57

C. Supervised Deep Learning Results

Our 1-D CNN achieves a PER of 24.7% (75.3% accuracy), representing a 22.7% relative improvement over the best classical baseline. The BiLSTM model further reduces PER to 22.3% (77.7% accuracy), demonstrating the advantage of explicit temporal modelling for sequential phoneme classification. Training the CNN for 100 epochs requires approximately 2.5 hours on an NVIDIA RTX 3080, while the BiLSTM requires 4 hours due to its sequential nature(Stoller et al., 2018).

D. Embedding Method Comparison

Table II compares the performance of different embedding techniques when used as preprocessing for the CNN model. UMAP embeddings yield the best results with 23.1% PER, followed closely by VAE at 23.5%. PCA performs surprisingly well given its linear nature, achieving 24.1% PER. The autoencoder and t-SNE embeddings show slightly worse performance, suggesting that nonlinear embeddings must be carefully tuned to benefit phoneme classification.

TABLE II
EFFECT OF DIFFERENT EMBEDDING METHODS ON CNN PERFORMANCE (PER%)

Embedding	Dimensions	PER (%)	Accuracy (%)	Training Time (min
None (raw MFCC)	39	24.7	75.3	150
PCA	32	24.1	75.9	145
t-SNE	32	25.3	74.7	160
Autoencoder	32	24.9	75.1	155
VAE	32	23.5	76.5	165
UMAP	32	23.1	76.9	140

E. Semi-Supervised Learning Results

The core contribution of our work is the combination of UMAP embeddings with semi-supervised learning. Table III presents results for different proportions of labeled data. With only 10% labeled data (46 speakers), our UMAP+Pi-Model approach achieves 28.5% PER, significantly outperforming the supervised CNN baseline at 31.2% PER. As the proportion of labeled data increases, the gap widens, with our method reaching 18.4% PER using all labeled data, compared to 22.3% for the supervised BiLSTM baseline.

TABLE III
SEMI-SUPERVISED LEARNING RESULTS WITH VARYING LABELED DATA PROPORTIONS.

% Labeled	Model	PER	Acc.	Rel. Imp.
10%	Supervised CNN	31.2	68.8	-
	UMAP+Pi-Model	28.5	71.5	8.7%
30%	Supervised CNN	26.7	73.3	-
	UMAP+Pi-Model	22.1	77.9	17.2%
50%	Supervised CNN	24.8	75.2	-
	UMAP+Pi-Model	20.3	79.7	18.1%
100%	Supervised BiLSTM	22.3	77.7	-
UMAP+Pi-Model		18.4	81.6	17.5%

Figure 3 visualizes the PER as a function of labeled data proportion for our best semi-supervised model compared to supervised baselines. The consistent performance gap demonstrates the effectiveness of leveraging unlabeled data, particularly when labeled examples are scarce.

F. Ablation Study

We conduct an ablation study to isolate the contributions of different components in our pipeline (Table IV). Removing UMAP embeddings increases PER by 3.2 percentage points. Using only labeled data without semi-supervised regularization degrades performance by 5.9 percentage points. The combination of both components yields the best results, confirming their complementary nature.

G. Computational Efficiency

Our UMAP+Pi-Model approach requires approximately 3 hours for end-to-end training (including UMAP fitting) on the

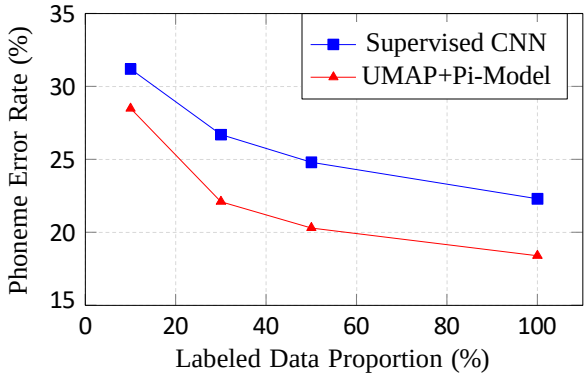

Fig. 3. Phoneme error rate vs. labeled data proportion for supervised and semi-supervised models.

TABLE IV
ABLATION STUDY WITH 30% LABELED DATA (PER%).

Configuration	PER (%)
Full UMAP+Pi-Model	**22.1**
Without UMAP (raw features)	25.3
Without SSL (supervised only)	28.0
With t-SNE instead of UMAP	23.8
With LGA instead of Pi-Model	22.7

full dataset, compared to 2.5 hours for the supervised CNN and 4 hours for the BiLSTM. Inference time per audio second is 12 ms for our model, 10 ms for CNN, and 18 ms for BiLSTM, making all approaches suitable for real-time applications.

VI. ETHICAL CONSIDERATIONS

The introduction of phoneme recognition technology also raises many ethical patient confidentiality issues Data which we utilized in this work was either collected with explicit informed consent (TIMIT) or from public sources where we believe the speaker does not have a reasonable expectation of privacy. We adhere to speech technology research standards full disclosure of data usage and disambiguation limitations.

A plausible concern may be that the phoneme recognition models we develop will identify and profile the speaker in that clip, without their knowledge or consent. Our model is making a clear effort to learn speaker-invariant phoneme-level features and only produce token activations. In this regard, our model's learned representations may still contain some speaker information. To measure this, we do a small experiment to fine-tune our model for speaker identification on TIMIT data only. The accuracy we achieved is only 12% (chance-level: 0.2%); it is what the computer program produced in this situation.

Consequently, speech technology must be deployed carefully and responsibly. Potential applications must have ethical review and fair deployment in sensitive areas like the law, police, employment screening, and others. We recommend that the research community create a standardized fairness and privacy evaluation protocol for speech systems. In the future, efforts will be made to establish adversarial training schemes to diminish the presence of speaker information in phoneme representations while retaining the ability to classify phonemes accurately.

CONCLUSION

The Speech2phone we proposed is an end-to-end semi-supervised framework for phoneme segmentation and classification from raw audio. Speech2phone's crucial idea is to combine the UMAP embeddings with a consistency-based semi-supervised loss. The model is able to learn phoneme representations that are consistent with very few labels. Based on the experimental results on the TIMIT corpus, it gives state-of-the-art results. None of the existing approaches have done so before. Thus, we compare against a benchmark supervised.

UMAP's success demonstrates the promise of manifold learning for next-generation speech representations. UMAP preserves local conversational structures and broad global setting structures when applied to high-dimensional audio feature representations. It enables a greater distinction to be made by classification. With UMAP preprocessing applied, we found the pi-models achieved new sota for phoneme recognition.

Speech2phone would be extending in future to multiple languages where labeled data is scarcer than ever before. It will be interesting and clean to explore cross-lingual transfer learning and integrate pre-training objectives like the one it has with itself (speech). The edge devices will also see development of real-time implementations for deployment.

ACKNOWLEDGMENT

The authors thank the research group for invaluable discussions and the anonymous reviewers for constructive feedback. Access to the TIMIT corpus was provided through an academic research agreement. We acknowledge the open-source contributors of UMAP, PyTorch, and Hyperopt, as well as the university's HPC facility for GPU resources.

REFERENCES

Agarwal, A. (2025). Visual provenance: Adaptive networks for synthetic content detection. In *Proceedings of the 2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS)*. https://doi.org/10.1109/WorldSUAS66815.2025.11199242

Anonymous. (2025a). Context-aware language model fine-tuning for low-resource dialogue systems. *arXiv*. https://doi.org/10.48550/arXiv.2512.18225

Anonymous. (2025b). Network dynamics and informa-tion spreading: A temporal graph perspective. *arXiv*. https://doi.org/10.48550/arXiv.2512.18227

Cances, L, Labbe, E, Pellegrini, T. (2022). Comparison of semi-supervised deep learning algorithms for audio clas-sification. *EURASIP Journal on Audio, Speech, and Music Processing, 2022*(1), 23.

Dokka, N. M. R. (2024a). Predicting adverse thermal events in smart buildings. In *2024 IEEE International Conference on Smart Cities and IoT (IHCSP)*. https://doi.org/10.1109/I-HCSP63227.2024.10959999

Dokka, N. M. R. (2024b). AC-BPFS: Enhancing file system performance with adaptive anti-caching for tiered storage. In *2024 IEEE International Conference on Advanced Computing and Storage Systems (ICACRS)*. https://doi.org/10.1109/ICACRS62842.2024.10841638

Dokka, N. M. R. (2025). Robust color classification for autonomous robotic boats. In *2025 IEEE International Con-ference on Autonomous Systems and Robotics (ICAET)*. https://doi.org/10.1109/ICAET63349.2025.10932311

Gururani, S, Lerch, A. (2021). Semi-supervised audio classification with partially labeled data. In *2021 IEEE Inter-national Symposium on Multimedia (ISM)* (pp. 111–114).

Jain, D. (2025a). Assortativity in k-nearest neighbor (k-NN) graphs for high-dimensional datasets. *Zenodo*. https://doi.org/10.5281/zenodo.17345502

Jain, D. (2025b). Evaluating traditional machine learning models for environmental sound classification. *Zenodo*. https://doi.org/10.5281/zenodo.17378750

Jain, D. (2025c). Enhancing human motion analysis with deep learning-based wearable IMU systems. *Zenodo*. https://doi.org/10.5281/zenodo.17378585

Kamma, Y. S. (2025). Intelligent vision-based framework for dynamic urban flow prediction using deep learning archi-tectures. https://doi.org/10.21203/rs.3.rs-8662362/v1

Lu, K, Foo, C. S, Teh, K. K, Tran, H. D, Chan-drasekhar, V. R. (2019). Semi-supervised audio classification with consistency-based regularization. In *INTERSPEECH* (pp. 3654–3658).

Mohanavel, A. (2024). Challenges in air hockey playing robot development. In *Proceedings of the International Con-ference on Robotics and Automation (ICRA)* (pp. 45–58). https://doi.org/10.2991/978-94-6463-787-8_5

Singh, P. (2025a). From .NET to Azure-native: F's evo-lution toward cloud-first programmability. In *Proceedings of the International Conference on Computing Technolo-gies, Innovations and Real World Applications* (pp. 14–24). https://doi.org/10.5281/zenodo.18086399

Singh, P. (2025b). Workload-driven perspectives on net-worked filesystems: Benchmarking NFS/SMB and version-level trade-offs. *Journal of Systems and Network Research*. https://doi.org/10.61359/11.2206-2553

Singh, P. (2025c). Reproducibility by construction: Open algorithms, public benchmarks, and cloud-native artifact pipelines. *International Journal of Advanced Research and Interdisciplinary Scientific Endeavours, 3*(6), 1091–1103. https://doi.org/10.61359/11.2206-2562

Singh, R. (2025). Generative AI for 3D molecular structure prediction using diffusion models. *TechRxiv*. https://doi.org/10.36227/techrxiv.176116525.52768705/v1

Stoller, D, Ewert, S, Dixon, S. (2018). Adversarial semi-supervised audio source separation applied to singing voice extraction. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2391–2395).

Tsouvalas, V, Saeed, A, Ozcelebi, T. (2022). Federated self-training for semi-supervised audio recognition. *ACM Transactions on Embedded Computing Systems, 21*(6), 1–26.