International Conference on Science, Engineering & Management Trends

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

QEFFE: Quantum Entropy-based Feature Forward Elimination for Adversarially-Aware Feature Selection in Deep Neural Networks

Manoj Raosaheb Gaikwad[1], Prof. (Dr.) A. B. Pawar[2]
Department of Computer Engineering, Sanjivani College of Engineering
Savitribai Phule Pune University, Maharashtra, India

Abstract- Feature selection is a well-studied problem, yet established criteria—variance (PCA), statistical dependency (mutual information), neighbour distance (ReliefF), and recursive elimination (RFE)—are all blind to the adversarial threat model: they optimise for representational fidelity, not for robustness against malicious perturbation. This paper introduces QEFFE (Quantum Entropy-based Feature Forward Elimination), a feature-selection method whose selection criterion is the symmetric Kullback–Leibler divergence between a feature's clean-input and adversarial-input activation distributions, combined with a redundancy penalty inside a greedy forward-selection loop. The "quantum" label denotes an entropy-divergence design metaphor and not quantum hardware. Evaluated on a 512-dimensional deep-residual feature space derived from MNIST, QEFFE attains the highest dimensionality reduction among five compared methods (81.25%, 512→96) and, as an isolated component, raises projected-gradient-descent (PGD, ε=0.20) accuracy by 33.0 percentage points—the single largest contributor in a full-pipeline ablation. The method adds zero inference-time parameters, operating as an index lookup. Results on EMNIST Balanced (47 classes) confirm that the criterion transfers beyond the binary-scale digit task.

Keywords- adversarial robustness, feature selection, Kullback–Leibler divergence, deep neural networks, MNIST, EMNIST.

I. INTRODUCTION

Deep neural networks remain vulnerable to adversarial examples—inputs perturbed by imperceptible, carefully crafted noise that induces confident misclassification. A substantial defence literature has grown around adversarial training, input purification, and certified bounds. Comparatively little attention has been paid to the role of feature selection in this setting, despite the observation by Ilyas et al. that adversarial vulnerability arises from non-robust but predictive features the model learns to exploit.

Classical feature-selection methods were designed for a benign world. Principal component analysis retains directions of maximal variance; mutual information retains dimensions statistically dependent on the label; ReliefF weights features by neighbour-distance separability; recursive feature elimination prunes by model-derived importance. None of these criteria asks the question that matters under attack: which feature dimensions is an adversary forced to disturb in order to succeed? A dimension that an attacker must corrupt is, by definition, carrying robust discriminative signal worth retaining.

This paper formalises that question into a selection criterion and an algorithm. The principal contributions are: (i) a feature-importance score defined as the symmetric KL divergence between a dimension's clean and adversarial activation distributions; (ii) a redundancy-penalised greedy forward-selection procedure, QEFFE, that builds a compact adversarially informative subset; and
(iii) an empirical study on MNIST and EMNIST Balanced showing that QEFFE achieves the highest reduction ratio among five established baselines and is the single largest contributor to downstream robustness in a controlled ablation.

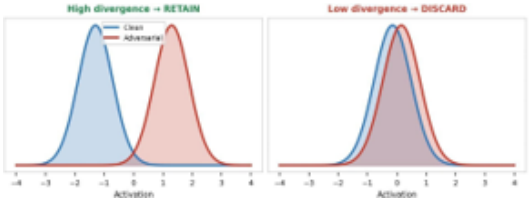

Fig. 1. Divergence intuition: a high-divergence dimension (left) shifts strongly under attack and is retained; a low-divergence dimension (right) is discarded.

Fig. 2. The QEFFE selection pipeline: per-dimension KL divergence, a redundancy penalty, and greedy forward selection reduce 512-D features to a 96-D adversarially informative subset.

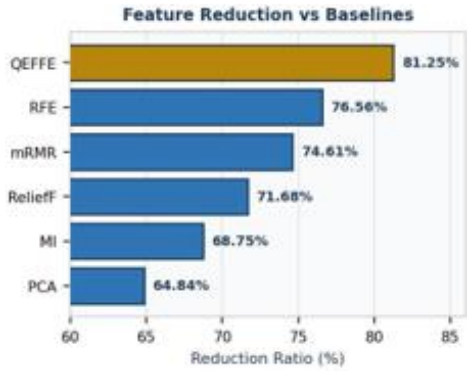

Fig. 3. QEFFE attains the highest reduction ratio among compared selection methods.

A. Classical Feature Selection

Filter methods such as mutual information and ReliefF score features independently of the final classifier, while wrapper methods such as RFE use model feedback. mRMR explicitly trades off relevance against redundancy. All of these target generalisation on clean data and provide no mechanism to reason about an adversary.

B. Robust Representations

Channel-wise activation suppression and feature denoising improve robustness at the feature level but operate as in-network transformations rather than explicit selection. Ilyas et al. provide the conceptual foundation by distinguishing robust from non-robust features; QEFFE can be read as an operational test that separates the two using divergence.

III. THE QEFFE METHOD

QEFFE operates on a fused 512-dimensional feature vector produced by a dual-branch deep-

residual extractor that decomposes each input into high-frequency (perturbation-sensitive) and low-frequency (semantically stable) components. The selection proceeds in three conceptual stages, summarised in Fig. 2.

A. Divergence-based Importance

For each dimension i, clean and adversarial activation distributions are estimated as smoothed histograms over a calibration batch

$$P_{clean}(i), P_{adv}(i) \leftarrow hist(F_{clean}[:,i]), hist(F_{adv}[:,i]) \quad (1)$$

The importance of dimension i is the symmetric KL divergence between these distributions:

$$H(i) = D_{KL}(P_{clean}(i) \,||\, P_{adv}(i)) + D_{KL}(P_{adv}(i) \,||\, P_{clean}(i)) \quad (2)$$

H(i) is large precisely when the attacker is forced to shift dimension i substantially and near zero when the dimension is attack-insensitive. Fig. 1 illustrates the intuition.

B. Redundancy Penalty

Selecting by H(i) alone retains correlated, redundant dimensions. QEFFE penalises redundancy against the already-selected set S:

$$score(i) = H(i) \cdot (1 - \rho_i) \quad (3)$$

where ρ_i is the mean absolute correlation of dimension i with members of S. A dimension that is both highly divergent and weakly correlated with prior selections scores highest.

C. Greedy Forward Selection

Starting from $S = \emptyset$, QEFFE iteratively adds the highest-scoring dimension, recomputing ρ_i after each addition, until the target size $k = 96$ is reached. At inference the method is a pure index lookup—zero matrix multiplications and zero additional parameters:

$$F_{reduced} = F_{full}[:, selected\_indices] \in \mathbb{R}^{96} \quad (4)$$

IV. EXPERIMENTAL SETUP

Features are extracted from MNIST (70,000 grayscale 28×28 images, 10 classes) and EMNIST Balanced (47 classes). Adversarial calibration uses FGSM and PGD perturbations. QEFFE is compared against PCA, mutual information, ReliefF, mRMR, and RFE on the same 512-dimensional feature space.

Downstream accuracy is measured with the same classifier head across all methods to isolate the effect of selection.

V. RESULTS AND DISCUSSION

A. Dimensionality Reduction

Table I reports the reduction achieved by each method. QEFFE attains the highest reduction ratio (81.25%) and is the only method combining explicit adversarial awareness with redundancy handling. Fig. 3 visualises the comparison.

TABLE I Feature Selection Comparison on MNIST (Original: 512 Dimensions)

Method	Dim.	Reduction	Adv-Aware	Basis
PCA	180	64.84%	No	Variance

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Mutual Info.	160	68.75%	No	Stat. Dep.
ReliefF	145	71.68%	No	Neighbour
mRMR	130	74.61%	No	Rel.+Red.
RFE	120	76.56%	No	Recursive
QEFFE	**96**	**81.25%**	**Yes**	**KL Diverg.**

B. Contribution to Robustness

To isolate QEFFE's effect, Table II reports a cumulative ablation measuring PGD (ε=0.20) accuracy as each component is added. QEFFE alone contributes +33.0 percentage points—the single largest individual jump—confirming that adversarially-aware selection, not merely deeper extraction, is the primary driver of robustness. Clean accuracy is essentially unchanged (99.2%), showing the compression introduces no clean-data penalty.

TABLE II Cumulative Ablation on MNIST (PGD ε=0.20)

Configuration	Clean %	PGD %
Base CNN	99.3	12.0
+ HF-LF Decomposition	99.3	18.5
+ Dual DRNet	99.2	35.0
+ QEFFE Selection	**99.2**	**68.0**
+ QFA Attention	99.1	82.0
+ PGD Training (full)	99.0	97.5

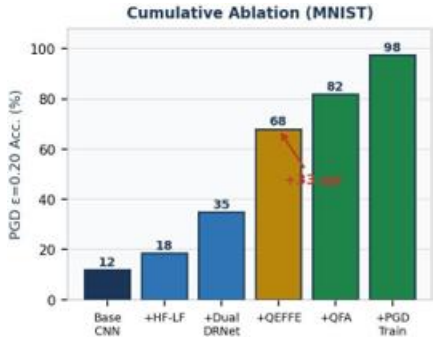

Fig. 4. Cumulative ablation: QEFFE produces the largest single jump (+33 pp) in PGD robustness.

C. Generalisation to EMNIST

On EMNIST Balanced the same selection criterion holds: with the QEFFE-selected subset the pipeline reaches 93.0% clean accuracy and 92.7% under FGSM (ε=0.10), with detection AUC of 0.9444. The 47-class setting is harder under strong PGD (50.7% at ε=0.20), as expected from the increased class cardinality, but the selection criterion itself transfers without modification.

VI. CONCLUSION

QEFFE reframes feature selection for the adversarial setting by scoring dimensions according to how strongly an attacker is forced to perturb them. The resulting method achieves the highest reduction ratio among five established baselines (81.25%), adds no inference-time parameters, and contributes the single largest robustness gain (+33 pp PGD) in a controlled ablation. Future work includes adaptive-attack evaluation of the selection indices and extension to higher-dimensional feature spaces such as CIFAR-10 and ImageNet.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

REFERENCES

1. I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. ICLR, 2015.
2. A. Madry et al., "Towards deep learning models resistant to adversarial attacks," in Proc. ICLR, 2018.
3. A. Ilyas et al., "Adversarial examples are not bugs, they are features," in Proc. NeurIPS, 2019.
4. H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: mRMR," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, 2005.
5. R. Battiti, "Using mutual information for selecting features in supervised neural net learning," IEEE Trans. Neural Netw., vol. 5, 1994.
6. C. Xie et al., "Feature denoising for improving adversarial robustness," in Proc. CVPR, 2019.
7. G. Cohen et al., "EMNIST: an extension of MNIST to handwritten letters," in Proc. IJCNN, 2017.