Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 06, No. 02, July-Dec 2026
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.62.84.97

Research Paper

Learning before memorising: How far does early stopping protect multilayer perceptrons from label noise?

Muntaser Syed[1*], Hidayath Ali Baig Mohammed[2]
[1*]Department of Electrical Engineering and Computer Science, Florida Institute of Technology, Melbourne, Florida, USA.
[2]Department of Computing and Information Sciences, College of Computing and Information Sciences, University of Technology and Applied Sciences, Sur, Sultanate of Oman.

Article Info

Article History:
Received: 19 June 2026
Revised: 18 August 2026
Accepted: 11 September 2026
Published: 22 October 2026

Keywords:
Label Noise
Early Stopping
Memorisation
Multilayer Perceptron
Robust Learning
Neural Network Training Dynamics

ABSTRACT

Supervised neural networks are trained on labels that are frequently wrong, and over-parameterised models are able to fit such errors exactly. This study quantifies, on a controlled benchmark, how much of the resulting damage is prevented by early stopping when the stopping signal itself comes from a validation split carrying the same label noise. Using the UCI handwritten digits dataset, training labels were corrupted with symmetric noise (0–60%) and with structured pair-flip noise (10–40%), and a multilayer perceptron (MLP) trained without regularisation, with strong weight decay (MLP-L2) or with early stopping (MLP-ES) was compared with logistic regression, a support vector machine, a random forest and k-nearest-neighbour classifiers over five random splits. Per-epoch tracking showed a clear learning-before-memorising phase under symmetric noise: with 40% corrupted labels, the MLP fitted 95% of the correctly labelled training examples by epoch 14 while reproducing only 2.2% of the corrupted ones, whereas memorisation of the latter exceeded 50% only after about 89 epochs. Consequently, clean test accuracy fell from a peak of 0.928 to 0.646 when training continued, while early stopping on noisy validation data retained 0.911. Across symmetric noise levels, MLP-ES improved on the unregularised MLP by 0.239 in accuracy on average (30 of 30 paired comparisons, Holm-adjusted $p < 0.001$) and reached 0.873 at 60% noise. Under pair-flip noise, however, the learning and memorising phases overlapped, the benefit shrank to 0.084, and MLP-ES was no better than weight decay. A random forest reproduced every corrupted label yet remained accurate, showing that memorisation alone does not determine harm. Early stopping is therefore an effective and nearly free defence against random label noise, but not a substitute for noise-aware methods when errors are systematic.

Corresponding Author:
Muntaser Syed

Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN)

Department of Electrical Engineering and Computer Science, Florida Institute of Technology, Melbourne, Florida, USA.
Email: msyed2011@my.fit.edu

Copyright © 2026 The Author(s). This is an open access article distributed under the Creative Commons Attribution License, (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Supervised learning assumes that training labels are correct, yet in practice they rarely are. Labels produced by human annotators, weak heuristics or automatic extraction contain errors whose rate commonly lies between a few per cent and several tens of per cent, and even curated benchmark test sets contain measurable proportions of mislabelled examples [1], [2]. The consequences of such class noise for classical learners have been studied for decades [3], [4], [5], [6], but the question has acquired new urgency with neural networks. [7] Showed that standard networks can fit completely random labels to zero training error, which means that nothing in the architecture itself prevents a network from memorising every wrong label it is given.

Memorisation, however, is not instantaneous. [8] Observed that networks trained by gradient descent first fit the simple, shared patterns of the correctly labelled data and only later memorise individual noisy examples. [9] Proved that for sufficiently over-parameterised networks, gradient descent with early stopping is provably robust to a constant fraction of corrupted labels, [10] reported that deep networks can tolerate large amounts of label noise under suitable training settings. These results suggest that early stopping [11], one of the oldest and cheapest regularisers, should be an effective defence against label noise. Much of the recent literature has nevertheless concentrated on more elaborate remedies such as loss correction [12], robust loss functions [13], [14], sample selection [15] and data cleaning [2]; see [16] for a survey.

Two practical questions remain insufficiently answered. First, in a real application the validation data used to decide when to stop are drawn from the same noisy source as the training data, so the stopping signal is itself corrupted. Does early stopping still find the right moment when it can only observe noisy validation accuracy? Second, the separation between learning and memorising may depend on the structure of the noise: errors that are spread uniformly across classes are statistically different from systematic confusions between specific classes, which are common in real annotation [6], [17]. How much protection does early stopping give against each kind of noise, and how does it compare with weight decay [18] and with classical models whose inductive biases differ?

This paper answers these questions with a controlled, reproducible benchmark on a public image dataset. The contributions are as follows:

A per-epoch measurement of the learning-before-memorising phase of a multilayer perceptron (MLP) under symmetric and pair-flip label noise, quantifying the width of the window in which clean patterns have been learned but corrupted labels have not yet been memorised.

Evidence that early stopping driven only by noisy validation accuracy recovers most of the attainable accuracy under symmetric noise, together with a paired statistical comparison against weight decay, an unregularised MLP and five classical classifiers across noise rates of up to 60%.

Evidence that the protective window largely disappears under structured pair-flip noise, where early stopping is no better than weight decay and simple neighbourhood averaging is more robust.

A dissociation between memorisation and harm: a random forest reproduces every corrupted training label yet generalises well, whereas an MLP that memorises the same labels degrades severely.

Section 2 reviews related work, Section 3 describes the experimental design, Section 4 reports the results, discusses their implications and limitations, and Section 5 concludes the paper.

2. RELATED WORK

2.1 Label Noise and Its Effects on Classifiers

Early theoretical work by Angluin and Laird [3] showed that learning remains possible under random classification noise, at the cost of more samples. Empirical studies found that class noise is generally more harmful than attribute noise [4] and that its impact varies strongly between learning algorithms [5]. Frénay and Verleysen [6] distinguished noise that is completely at random, noise that depends on the true class, and noise that depends on both the class and the input, and reviewed robust, cleansing and noise-modelling approaches. [19] Derived unbiased estimators for class-conditional noise, and [12] proposed forward and backward loss corrections for deep networks based on an estimated noise transition matrix.

2.2 Memorisation in Neural Networks

The ability of neural networks to interpolate arbitrary labels [7] challenged classical explanations of generalisation. [8] Showed that real data are learned before noise and that regularisation slows memorisation of noisy examples more than learning of clean ones. Related work on interpolation shows that fitting training data exactly need not harm generalisation, depending on how the fitted function behaves between training points [20], [21]; [22] argued that random forests and AdaBoost succeed precisely as self-averaging interpolating classifiers. These results imply that the relevant question is not whether a model memorises noisy labels, but how the memorised labels influence predictions on new data.

2.3 Defences against Label Noise

Robust losses such as mean absolute error [13] and generalised cross-entropy [14] bound the influence of confidently wrong labels. Co-teaching [15] trains two networks that select small-loss examples for each other, exploiting the observation that clean examples are learned first. Confident learning [2] estimates the joint distribution of noisy and true labels in order to find and remove errors. Classical regularisers, namely weight decay [18], dropout [23] and early stopping [11], act more indirectly by limiting model capacity or training time. [9] Provided theoretical support for early stopping, but their guarantees assume idealised stopping times, whereas practitioners must stop on the basis of noisy validation data. [17] Showed on real human annotations that noise in practice is structured rather than uniform, which motivates evaluating defences under both regimes.

3. METHODOLOGY

The experimental design is summarised in Figure 1. All corruption was applied to training labels only; test labels were always the original, clean labels, so that test accuracy measures what the model has learned about the true task.

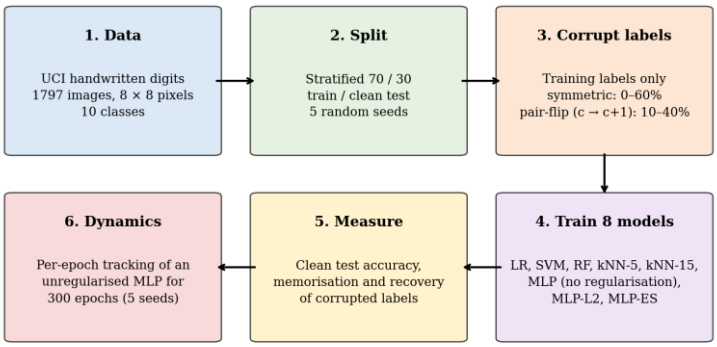


Figure 1. Overview of the Experimental Design. Label Corruption is applied only to the Training Data, and all Models are evaluated on Clean Test Labels

3.1 Dataset

The UCI optical recognition of handwritten digits dataset [24], in the version distributed with scikit-learn [25], contains 1797 grey-scale images of handwritten digits (0–9) represented on an 8 × 8 grid with 17 intensity levels, giving 64 features. The classes are nearly balanced (174–183 images per class). Pixel values were divided by 16 so that all features lie in [0, 1]. For each of five random seeds, the data were split into a stratified training set (70%, 1257 images) and a clean test set (30%, 540 images).

3.2 Label-Noise Models

Two noise models were applied to the training labels. Under symmetric noise at rate η, a randomly chosen fraction η of training examples had their label replaced by one of the nine other classes with equal probability; $\eta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$. Under pair-flip noise at rate η, a random fraction η of examples of class c were relabelled as class $(c + 1) \bmod 10$; $\eta \in \{0.1, 0.2, 0.3, 0.4\}$. Pair-flip noise represents systematic confusion between two specific classes and is a standard structured-noise benchmark [15]. At $\eta = 0.4$ it leaves the correct class only a narrow majority within each class, making it the hardest condition studied. The corruption transition matrix for symmetric noise is [1]where $\tilde{y}$ denotes the observed (possibly corrupted) label and y the true label.

3.3 Models

Eight classifiers were compared (Table 1). Three variants of the same MLP architecture [26], with two hidden layers of 256 and 128 ReLU units trained with Adam [27], isolate the effect of regularisation: an unregularised network trained for a fixed 500 epochs (MLP), a network with strong weight decay (MLP-L2, $\alpha = 1.0$) and a network with early stopping (MLP-ES). MLP-ES holds out 15% of the (noisy) training data as a validation split, monitors accuracy on this split after every epoch and stops when it has not improved for 20 consecutive epochs, restoring the best weights; the validation labels are corrupted in exactly the same way as the training labels. The classical baselines were multinomial logistic regression (LR) [28], a support vector machine with radial-basis kernel (SVM) [29], a random forest (RF) [30] and k-nearest-neighbour classifiers [31] with $k = 5$ and $k = 15$. Hyperparameters were fixed a priori (Table 1) and not tuned on noisy data, because tuning on corrupted validation labels would confound the comparison.

Table 1. Classifiers and their Fixed Configurations

Model	Description	Configuration
LR	Multinomial logistic regression	L2 penalty, C = 1.0, L-BFGS, max. 3000 iterations
SVM	Support vector machine	RBF kernel, C = 10, γ = "scale"
RF	Random forest	300 trees, $\sqrt{p}$ features per split, fully grown trees
kNN-5	k-nearest neighbours	k = 5, Euclidean distance, uniform weights
kNN-15	k-nearest neighbours	k = 15, Euclidean distance, uniform weights

MLP	Unregularised multilayer perceptron	(256, 128) ReLU, Adam (lr 10^{-3}), $\alpha = 10^{-5}$, 500 epochs, no early stopping
MLP-L2	MLP with weight decay	As MLP, $\alpha = 1.0$
MLP-ES	MLP with early stopping	As MLP, 15% noisy validation split, patience 20 epochs, best weights restored

3.4 Metrics

Performance was measured by accuracy on the clean test set. To separate memorisation from learning, two quantities were computed on the corrupted subset B of the training data, that is, the examples whose observed label differs from the true label. The memorisation rate is the proportion of these examples for which the fitted model predicts the corrupted label, and the recovery rate is the proportion for which it predicts the true label: [2]

A model that has learned the task without absorbing the noise has low memorisation and high recovery; a model that has interpolated the noisy training set has a memorisation rate of one.

3.5 Training-Dynamics Experiment

To observe learning and memorisation over time, the unregularised MLP was trained for 300 epochs by repeated single-epoch updates (mini-batches of 200) under four conditions: clean labels, 20% and 40% symmetric noise, and 30% pair-flip noise, with five seeds each. After every epoch we recorded clean test accuracy, the fit on correctly labelled training examples, the memorisation rate on corrupted training examples, and accuracy on a 15% validation split scored against both its noisy and its true labels. This allowed us to compare the epoch that would be selected by noisy validation accuracy with the epoch of peak clean test accuracy, a quantity that is unobservable in practice.

3.6 Statistical Analysis and Implementation

Results are reported as mean ± standard deviation over five seeds. Overall differences among the eight classifiers were tested with the Friedman test [32] on the mean accuracy over symmetric noise rates 0–50% per seed. Selected pairwise comparisons were tested with the Wilcoxon signed-rank test [33], pairing results by seed and noise rate (30 pairs for symmetric noise, 20 for pair-flip noise), with Holm correction over all 14 comparisons [34]. Because the same seed contributes several pairs, these tests are interpreted together with effect sizes and win counts [35]. All experiments were implemented in Python 3 with scikit-learn 1.8.0 [25] on a single CPU core with fixed random seeds, and the code is available from the corresponding author.

4. RESULTS AND DISCUSSION

4.1 Accuracy under Symmetric Label Noise

Table 2 and Figure 2(a) report clean test accuracy as the proportion of corrupted training labels increases. With clean labels, all classifiers reached 0.965–0.990. The unregularised MLP was the most noise-sensitive model: its accuracy fell almost linearly, to 0.639 at 40% and 0.446 at 60% noise. Early stopping changed this picture completely. MLP-ES lost only 0.042 of accuracy between 0% and 40% noise (0.966 to 0.924) and retained 0.905 at 50% and 0.873 at 60%, the highest value of all eight models at 60%. Weight decay also helped substantially up to about 40% (0.890) but degraded faster thereafter (0.704 at 60%). Among the classical models, kNN-15 was the most robust up to 50% noise (0.933), followed by RF (0.917), whereas kNN-5 and SVM degraded markedly.

Table 2. Clean Test Accuracy under Symmetric Label Noise (Mean ± SD over Five Seeds)

Noise Rate	LR	SVM	RF	KNN-5	KNN-15	MLP	MLP-L2	MLP-ES
0%	0.965 ± 0.003	0.990 ± 0.004	0.973 ± 0.006	0.983 ± 0.003	0.973 ± 0.010	0.977 ± 0.005	0.970 ± 0.006	0.966 ± 0.005

10%	0.949 ± 0.007	0.972 ± 0.002	0.971 ± 0.004	0.977 ± 0.004	0.971 ± 0.011	0.911 ± 0.012	0.967 ± 0.003	0.959 ± 0.006
20%	0.937 ± 0.011	0.953 ± 0.008	0.962 ± 0.006	0.965 ± 0.006	0.973 ± 0.006	0.831 ± 0.010	0.958 ± 0.010	0.948 ± 0.010
30%	0.917 ± 0.005	0.922 ± 0.008	0.958 ± 0.006	0.930 ± 0.009	0.970 ± 0.005	0.757 ± 0.017	0.935 ± 0.009	0.936 ± 0.007
40%	0.897 ± 0.009	0.877 ± 0.009	0.941 ± 0.007	0.854 ± 0.024	0.968 ± 0.007	0.639 ± 0.019	0.890 ± 0.017	0.924 ± 0.009
50%	0.853 ± 0.013	0.837 ± 0.017	0.917 ± 0.010	0.746 ± 0.037	0.933 ± 0.016	0.530 ± 0.027	0.838 ± 0.030	0.905 ± 0.020
60%	0.800 ± 0.035	0.740 ± 0.031	0.841 ± 0.032	0.620 ± 0.032	0.862 ± 0.025	0.446 ± 0.013	0.704 ± 0.024	0.873 ± 0.036

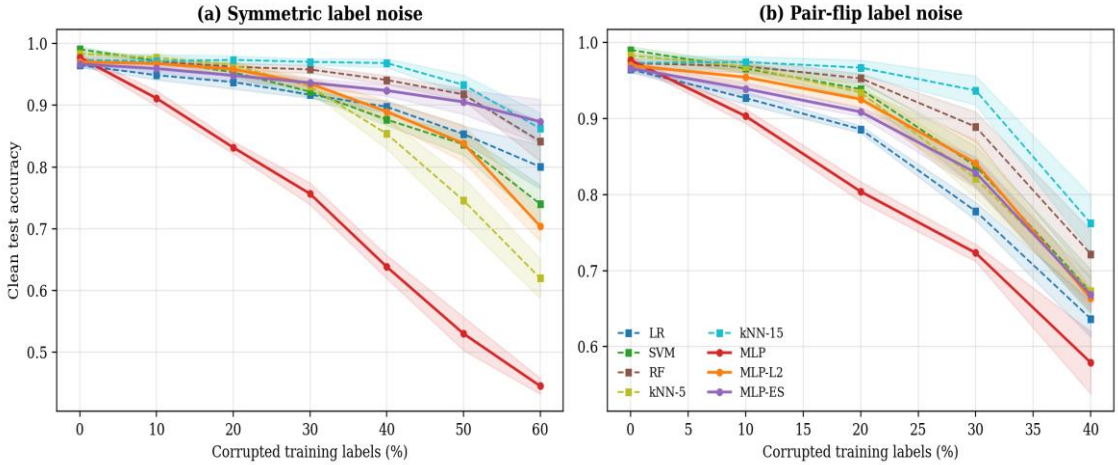


Figure 2. Clean Test Accuracy as a Function of the Proportion of Corrupted Training Labels under (a) Symmetric and (b) Pair-Flip Noise. Shaded Bands Show ±1 SD over Five Seeds; Solid Lines Denote the Three MLP Variants

Averaged over symmetric noise rates from 0% to 50%, the ranking was kNN-15 (0.965), RF (0.954), MLP-ES (0.940), MLP-L2 (0.926), SVM (0.925), LR (0.920), kNN-5 (0.909) and MLP (0.774); the differences among classifiers were significant (Friedman χ^2 = 33.67, df = 7, p = 2.0 × 10−5). The paired comparisons in Table 6 confirm that early stopping improved on the unregularised MLP in all 30 paired comparisons by 0.239 on average, and outperformed weight decay by 0.042 (20 of 30 comparisons, adjusted p = 0.013). The difference between RF and MLP-ES was small and not significant (0.007, p = 0.057).

4.2 Memorisation versus Harm

Table 3 shows how much of the corrupted training data each model reproduced. Early stopping kept memorisation of corrupted labels at or below 5.3% at every symmetric noise level, and the model recovered the true label for 83–96% of corrupted examples. Weight decay allowed memorisation to rise steadily with the noise rate, from 4.3% to 32.3%. The unregularised MLP reproduced 100% of corrupted labels at every noise level, as expected for a network trained to interpolation [7].

Table 3. Memorisation Rate (%) of Corrupted Training Labels under Symmetric Noise: the Percentage of Corrupted Training Examples for Which the Fitted Model Predicts the Corrupted Label (Mean over Five Seeds)

Noise Rate	LR	SVM	RF	KNN-5	KNN-15	MLP	MLP-L2	MLP-ES
10%	2.4	37.9	100.0	1.6	0.2	100.0	4.3	3.3

20%	2.0	37.8	100.0	2.7	0.3	100.0	5.7	1.4
30%	2.5	41.4	100.0	6.4	0.5	100.0	11.9	2.9
40%	2.8	42.0	100.0	11.0	0.7	100.0	15.4	2.5
50%	4.0	44.5	100.0	15.6	2.0	100.0	20.3	2.2
60%	6.4	53.6	100.0	22.9	5.0	100.0	32.3	5.3

RF reproduces all training labels because each training point is contained in the bootstrap samples of most trees and fully grown trees fit their bootstrap samples exactly.

The random forest, however, also reproduced every corrupted training label, yet it was the second most accurate model on clean test data. Figure 3 makes this dissociation explicit: at 40% symmetric noise, RF and MLP both memorised 100% of the corrupted labels, but their test accuracies were 0.941 and 0.639, respectively. Memorising a wrong label means that the model predicts that label at the exact training point; whether this harms predictions elsewhere depends on how far the influence of the point extends. In a random forest, each training point affects only the leaves that contain it, and its vote is diluted by the many trees built on bootstrap samples that exclude it [22]. In the MLP, fitting a corrupted label requires adjusting shared weights, which distorts the decision function around that point and for neighbouring inputs. Harm is thus a property of how a model interpolates, not of whether it interpolates [20], [21].

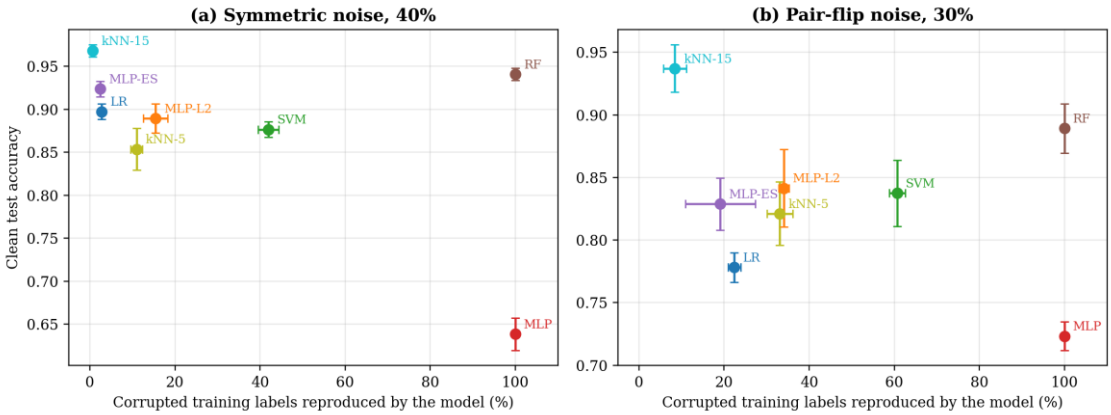


Figure 3. Memorisation of Corrupted Training Labels (X-Axis) Versus Clean Test Accuracy (Y-Axis) at (a) 40% Symmetric and (b) 30% Pair-Flip Noise. Error Bars Show ±1 SD over Five Seeds

4.3 Training Dynamics: The Learning-Before-Memorising Window

Figure 4 and Table 4 show the per-epoch behaviour of the unregularised MLP. Under symmetric noise, the network first fitted the correctly labelled training examples and only much later the corrupted ones. With 40% symmetric noise, it fitted 95% of the correctly labelled examples by epoch 14.0 ± 2.3, when it reproduced only 2.2 ± 0.9% of the corrupted labels; memorisation of the corrupted labels crossed 50% only at epoch 88.6 ± 6.2. Clean test accuracy peaked at 0.928 ± 0.011 around epoch 17 and then declined as memorisation set in, to 0.646 ± 0.015 at epoch 300. The learning-before-memorising window therefore spanned roughly 75 epochs, and the post-peak decline of 0.28 in accuracy was entirely attributable to memorisation.

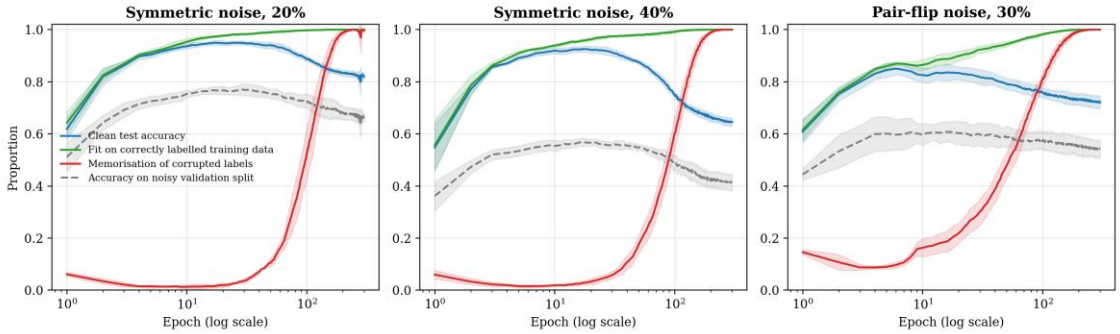


Figure 4. Per-Epoch Training Dynamics of the Unregularised MLP (Mean ± SD over Five Seeds; Logarithmic Epoch Axis). Memorisation of Corrupted Labels (Red) Begins Only after the Correctly Labelled Data Have Been Fitted (Green) under Symmetric Noise, but the Two Processes Overlap under Pair-Flip Noise. The Dashed Grey Curve is the Accuracy on a Noisy Validation Split, the Only Signal Available to Early Stopping in Practice

Table 4. Summary of Training Dynamics of the Unregularised MLP over 300 Epochs (Mean ± SD over Five Seeds)

Condition	Peak Test Accuracy	Epoch of Peak	Accuracy at Epoch 300	Selected by Noisy Validation	Selected by Clean Validation*	Epoch at 95% Clean Fit	Memorisation at That Epoch (%)	Epoch at 50% Memorisation
Clean labels	0.981 ± 0.003	142 ± 69	0.978 ± 0.002	0.979 ± 0.004	0.979 ± 0.004	8.8 ± 1.2	–	–
Symmetric 20%	0.953 ± 0.009	21.8 ± 6.7	0.820 ± 0.014	0.948 ± 0.008	0.948 ± 0.009	9.8 ± 2.2	1.6 ± 0.7	98.6 ± 10.3
Symmetric 40%	0.928 ± 0.011	16.6 ± 5.5	0.646 ± 0.015	0.911 ± 0.019	0.923 ± 0.012	14.0 ± 2.3	2.2 ± 0.9	88.6 ± 6.2
Pair-flip 30%	0.854 ± 0.026	7.8 ± 4.1	0.721 ± 0.025	0.820 ± 0.032	0.838 ± 0.026	51.4 ± 10.1	47.1 ± 11.9	53.4 ± 7.1

*Oracle selection using the true labels of the validation split, which are unavailable in practice; shown only as an upper reference for noisy-validation early stopping.

The practical question is whether this window can be found without clean labels. It can, to a large extent. Under symmetric noise, a corrupted label is spread evenly over the nine wrong classes, so a model that predicts the true class more often also agrees more often with the noisy validation label; noisy validation accuracy is therefore an increasing function of clean accuracy and peaks at nearly the same epoch (Figure 4, dashed curves). Selecting the epoch with the highest noisy validation accuracy yielded 0.948 at 20% noise, identical to oracle selection with clean validation labels, and 0.911 at 40% noise, compared with 0.923 for the oracle and a peak of 0.928 on the test set.

Pair-flip noise behaved differently. Because a fraction of each class is consistently relabelled as a specific neighbouring class, the corrupted labels form a coherent, learnable pattern rather than random scatter, so the network began to fit them almost immediately. At the epoch where 95% of the correctly labelled examples had been fitted (epoch 51.4 ± 10.1), 47.1 ± 11.9% of the corrupted labels had already been memorised, and 50% memorisation was reached shortly afterwards (epoch 53.4). Test accuracy peaked early, at 0.854 around epoch 8, and noisy-validation selection recovered 0.820 with a highly variable stopping epoch (26.4 ± 35.4). The learning-before-memorising window, which underlies the success of early stopping, had essentially closed.

4.4 Accuracy under Pair-Flip Noise

The consequences are visible in Table 5 and Figure 2(b). All models degraded far more under pair-flip noise than under symmetric noise at the same rate; at 40%, accuracy ranged only from 0.579 (MLP) to 0.762 (kNN-15). Early stopping still improved on the unregularised MLP (0.084 on average, 20 of 20 comparisons, adjusted $p < 0.001$), but it no longer outperformed weight decay: MLP-L2 was slightly better than MLP-ES (by 0.010, winning 14 of 20 comparisons, adjusted $p = 0.018$). kNN-15 was the most robust model under structured noise, exceeding MLP-ES by 0.074 on average and RF by a smaller margin, while LR degraded fastest among the classical models.

Table 5. Clean Test Accuracy under Pair-Flip Label Noise (Mean ± SD over Five Seeds)

Noise Rate	LR	SVM	RF	kNN-5	kNN-15	MLP	MLP-L2	MLP-ES
10%	0.927 ± 0.008	0.966 ± 0.006	0.969 ± 0.008	0.969 ± 0.006	0.974 ± 0.008	0.903 ± 0.009	0.954 ± 0.010	0.939 ± 0.008
20%	0.886 ± 0.005	0.939 ± 0.011	0.953 ± 0.006	0.934 ± 0.008	0.967 ± 0.010	0.804 ± 0.013	0.926 ± 0.012	0.909 ± 0.006
30%	0.778 ± 0.012	0.837 ± 0.026	0.889 ± 0.020	0.821 ± 0.025	0.937 ± 0.019	0.723 ± 0.011	0.841 ± 0.031	0.829 ± 0.021
40%	0.637 ± 0.024	0.672 ± 0.029	0.722 ± 0.031	0.673 ± 0.017	0.762 ± 0.037	0.579 ± 0.041	0.663 ± 0.014	0.668 ± 0.022

Table 6. Paired Comparisons of Clean Test Accuracy (Wilcoxon Signed-Rank Test, Pairs Formed by Seed and Noise Rate; Holm Correction over All 14 Tests)

Noise Type	Comparison (A vs B)	Mean Difference (A − B)	Wins of A	Adjusted P
Symmetric (10–60%)	MLP-ES vs MLP	+0.239	30 / 30	< 0.001
	MLP-L2 vs MLP	+0.196	30 / 30	< 0.001
	MLP-ES vs MLP-L2	+0.042	20 / 30	0.013
	RF vs MLP	+0.246	30 / 30	< 0.001
	kNN-15 vs kNN-5	+0.098	26 / 30	< 0.001
	kNN-15 vs MLP-ES	+0.022	26 / 30	< 0.001
	RF vs MLP-ES	+0.007	23 / 30	0.057
Pair-flip (10–40%)	MLP-ES vs MLP	+0.084	20 / 20	< 0.001
	MLP-L2 vs MLP	+0.094	20 / 20	< 0.001
	MLP-ES vs MLP-L2	−0.010	6 / 20	0.018
	RF vs MLP	+0.131	20 / 20	< 0.001
	kNN-15 vs kNN-5	+0.061	19 / 20	< 0.001
	kNN-15 vs MLP-ES	+0.074	20 / 20	< 0.001
	RF vs MLP-ES	+0.047	20 / 20	< 0.001

Early stopping was also inexpensive. MLP-ES stopped after 34–63 epochs on average across conditions and required 0.69 s per fit, compared with 8.5 s for the 500-epoch MLP and 8.4 s for MLP-L2, so it was both the most accurate MLP variant under symmetric noise and the cheapest to train.

4.5 Discussion

4.5.1 Principal Findings

Under symmetric label noise, a small MLP learns the true task well before it memorises the wrong labels, and the moment at which to stop can be identified from validation data that are themselves noisy. Early stopping on noisy validation accuracy recovered most of the peak accuracy, kept memorisation of corrupted labels below about 5% at noise rates of up to 60%, and converted the least robust classifier of the study into one of the most robust. This confirms, on a practical benchmark and without clean validation labels, the empirical observations [8] and the theoretical predictions [9].

The protection, however, depends on the structure of the noise. Under pair-flip noise the corrupted labels are consistent enough to be learned alongside the true patterns, the window between learning and memorising closes, and early stopping loses its advantage over weight decay. This agrees with the observation that real annotation noise is class-dependent and systematic [6], [17] and implies that results obtained under symmetric noise, which is the most common synthetic benchmark, can overstate the value of timing-based defences.

4.5.2 Why Noisy Validation Works under Symmetric Noise

For symmetric noise with rate η and $K = 10$ classes, a prediction that equals the true label agrees with the observed label with probability $1 - \eta$, and a wrong prediction agrees with probability at most $\eta/9$. Expected noisy validation accuracy is therefore an increasing affine function of clean accuracy whenever $\eta < 0.9$, so both are maximised at the same point in expectation, and only sampling variability of the finite validation split separates the selected epoch from the optimal one. The slightly larger gap at 40% noise (0.911 selected versus 0.923 oracle) reflects this variability, because the effective signal shrinks as η grows. Under pair-flip noise the argument fails: predicting the flipped class $c + 1$ for an example of class c is rewarded by the corrupted validation labels, so noisy validation accuracy partly rewards memorisation, and the selected epoch becomes erratic.

4.5.3 Implications for Practice

Four recommendations follow from these results. First, early stopping on a held-out split of the noisy data should be the default for neural networks trained on labels of uncertain quality; it is nearly free and, under random noise, highly effective. Second, when there is reason to believe that label errors are systematic, for example confusions between visually or semantically similar classes, early stopping should not be relied upon alone, and noise-aware methods such as loss correction [12], sample selection [15] or confident learning [2] should be considered. Third, simple neighbourhood-averaging models such as kNN with a moderately large k and random forests are strong and robust baselines in the presence of label noise and should be included in comparisons. Fourth, a model that fits noisy training labels exactly is not necessarily harmed by them; robustness should be judged on clean held-out data, not on training-set memorisation.

4.5.4 Limitations

The study used a single, low-dimensional image dataset and small fully connected networks; convolutional and larger networks on high-resolution data have a different capacity and inductive bias, and the width of the learning-before-memorising window may differ. Only two synthetic noise models were examined; real annotation noise is often instance-dependent [6], [17]. Hyperparameters were fixed rather than tuned, which makes the comparison fair with respect to the absence of clean labels but may not represent the best achievable performance of each model. Five seeds provide stable estimates but limited statistical power, and the pooled paired tests treat results at different noise rates of the same seed as separate pairs; the conclusions therefore rest primarily on consistent effect sizes and win counts. Finally, the oracle comparisons with clean validation labels serve only as reference points and are not available in practice.

5. CONCLUSION

This study measured how far early stopping protects multilayer perceptrons from label noise when the stopping decision is made on equally noisy validation data. Under symmetric noise, the network learned the correct patterns tens of epochs before it memorised the corrupted labels; early stopping exploited this window, kept memorisation of wrong labels near 2–5%, preserved 0.924 accuracy with 40% corrupted labels and 0.873 with 60%, and improved on an unregularised network by 0.239 on average at a fraction of the training cost. Under structured pair-flip noise, learning and memorising overlapped, and early stopping offered no advantage over weight decay while neighbourhood-based classifiers were more robust. A random forest that memorised every corrupted label nevertheless generalised well, showing that the harm of label noise depends on how a model interpolates rather than on whether it memorises. Early stopping should thus be treated as a strong default against random label noise but not as a substitute for noise-aware learning when label errors are systematic. Future work will extend the analysis to convolutional networks, larger image and text datasets, and instance-dependent and real annotation noise.

Acknowledgments

The authors thank the creators of the UCI optical recognition of handwritten digits dataset and the developers of the open-source scientific Python ecosystem, as well as the researchers whose work is cited in this article.

Funding Information

This research has received no external funding.

Author Contribution Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Muntaser Syed	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Hidayath Ali Baig Mohammed	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

No conflict of interest.

Informed Consent

Not Applicable.

Ethical Approval

Not Applicable.

Data Availability

This study draws on publicly available data: the UCI optical recognition of handwritten digits dataset, as distributed with scikit-learn. No new datasets were created. The analysis code is available from the corresponding author on reasonable request.

REFERENCES

[1] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," in Proc. NeurIPS Datasets and Benchmarks Track, 2021. doi.org/10.48550/arXiv.2103.14749

[2] C. G. Northcutt, L. Jiang, and I. L. Chuang, "Confident learning: Estimating uncertainty in dataset labels," Journal of Artificial Intelligence Research, vol. 70, pp. 1373–1411, 2021. doi.org/10.1613/jair.1.12125

[3] D. Angluin and P. Laird, "Learning from noisy examples," Machine Learning, vol. 2, no. 4, pp. 343–370, 1988. doi.org/10.1007/BF00116829

[4] X. Zhu and X. Wu, "Class noise vs. attribute noise: A quantitative study," Artificial Intelligence Review, vol. 22, no. 3, pp. 177–210, 2004. doi.org/10.1007/s10462-004-0751-8

[5] D. F. Nettleton, A. Orriols-Puig, and A. Fornells, "A study of the effect of different types of noise on the precision of supervised learning techniques," Artificial Intelligence Review, vol. 33, no. 4, pp. 275–306, 2010. doi.org/10.1007/s10462-010-9156-z

[6] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 5, pp. 845–869, 2014. doi.org/10.1109/TNNLS.2013.2292894

[7] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," Communications of the ACM, vol. 64, no. 3, pp. 107–115, 2021. doi.org/10.1145/3446776

[8] D. Arpit et al., "A closer look at memorization in deep networks," in Proc. 34th Int. Conf. Machine Learning, PMLR, vol. 70, 2017, pp. 233–242. proceedings.mlr.press/v70/arpit17a.html

[9] M. Li, M. Soltanolkotabi, and S. Oymak, "Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks," in Proc. 23rd Int. Conf. Artificial Intelligence and Statistics, PMLR, vol. 108, 2020, pp. 4313–4324. proceedings.mlr.press/v108/li20j.html

[10] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," arXiv preprint arXiv:1705.10694, 2017. doi.org/10.48550/arXiv.1705.10694

[11] L. Prechelt, "Early stopping — but when?," in Neural Networks: Tricks of the Trade, 2nd ed., Lecture Notes in Computer Science, vol. 7700. Berlin: Springer, 2012, pp. 53–67. doi.org/10.1007/978-3-642-35289-8_5

[12] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2017, pp. 1944–1952. doi.org/10.1109/CVPR.2017.240

[13] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in Proc. 31st AAAI Conf. Artificial Intelligence, 2017, pp. 1919–1925. doi.org/10.1609/aaai.v31i1.10894

[14] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in Advances in Neural Information Processing Systems 31, 2018, pp. 8778–8788. doi.org/10.48550/arXiv.1805.07836

[15] B. Han et al., "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in Advances in Neural Information Processing Systems 31, 2018, pp. 8527–8537. doi.org/10.48550/arXiv.1804.06872

[16] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 11, pp. 8135–8153, 2023. doi.org/10.1109/TNNLS.2022.3152527

[17] J. Wei, Z. Zhu, H. Cheng, T. Liu, G. Niu, and Y. Liu, "Learning with noisy labels revisited: A study using real-world human annotations," in Proc. 10th Int. Conf. Learning Representations, 2022. doi.org/10.48550/arXiv.2110.12088

[18] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in Advances in Neural Information Processing Systems 4, 1991, pp. 950–957. papers.nips.cc/paper/563-a-simple-weight-decay-can-improve-generalization

[19] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari, "Learning with noisy labels," in Advances in Neural Information Processing Systems 26, 2013, pp. 1196–1204. papers.nips.cc/paper/5073-learning-with-noisy-labels
[20] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," Proceedings of the National Academy of Sciences, vol. 116, no. 32, pp. 15849–15854, 2019. doi.org/10.1073/pnas.1903070116
[21] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," Proceedings of the National Academy of Sciences, vol. 117, no. 48, pp. 30063–30070, 2020. doi.org/10.1073/pnas.1907378117
[22] A. J. Wyner, M. Olson, J. Bleich, and D. Mease, "Explaining the success of AdaBoost and random forests as interpolating classifiers," Journal of Machine Learning Research, vol. 18, no. 48, pp. 1–33, 2017. jmlr.org/papers/v18/15-240.html
[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, no. 56, pp. 1929–1958, 2014. jmlr.org/papers/v15/srivastava14a.html
[24] E. Alpaydin and C. Kaynak, "Optical recognition of handwritten digits," UCI Machine Learning Repository, 1998. archive.ics.uci.edu/dataset/80/optical+recognition+of+handwritten+digits
[25] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. jmlr.org/papers/v12/pedregosa11a.html
[26] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536, 1986. doi.org/10.1038/323533a0
[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. 3rd Int. Conf. Learning Representations, 2015. doi.org/10.48550/arXiv.1412.6980
[28] D. R. Cox, "The regression analysis of binary sequences," Journal of the Royal Statistical Society: Series B (Methodological), vol. 20, no. 2, pp. 215–242, 1958. doi.org/10.1111/j.2517-6161.1958.tb00292.x
[29] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995. doi.org/10.1007/BF00994018
[30] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi.org/10.1023/A:1010933404324
[31] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967. doi.org/10.1109/TIT.1967.1053964
[32] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," Journal of the American Statistical Association, vol. 32, no. 200, pp. 675–701, 1937. doi.org/10.1080/01621459.1937.10503522
[33] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bulletin, vol. 1, no. 6, pp. 80–83, 1945. doi.org/10.2307/3001968
[34] S. Holm, "A simple sequentially rejective multiple test procedure," Scandinavian Journal of Statistics, vol. 6, no. 2, pp. 65–70, 1979. www.jstor.org/stable/4615733
[35] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006. jmlr.org/papers/v7/demsar06a.html

How to Cite: Muntaser Syed, Hidayath Ali Baig Mohammed (2026). Learning before memorising: How far does early stopping protect multilayer perceptrons from label noise?. Journal of Artificial Intelligence, Machine Learning and Neural Network, 6(2), 84-97. https://doi.org/10.55529/jaimlnn.62.84.97

BIOGRAPHIES OF AUTHORS

	**Muntaser Syed**, is a researcher in the Department of Electrical Engineering and Computer Science at the Florida Institute of Technology, Melbourne, Florida, USA. His work covers deep learning, transfer learning and natural language processing, including topic modelling and text classification. He is interested in how learning systems behave under imperfect supervision and adversarial input, and in the empirical study of their failure modes. Email: msyed2011@my.fit.edu
	**Hidayath Ali Baig Mohammed**, is a member of the Department of Computing and Information Sciences, College of Computing and Information Sciences, University of Technology and Applied Sciences, Sur, Sultanate of Oman. His interests lie in applied artificial intelligence and data analytics, with attention to the quality of the data on which learning systems are trained. He works on the practical application of machine learning methods and on their evaluation in teaching and applied settings. Email: hidayathalibaig@gmail.com