International Journal of Research and Development in Engineering Sciences
ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/
Regular Issue, Vol. 8, Issue. 6 (November – December) , 2026 , Pages: 88 – 96
Received: 01 May 2026 ; *Accepted:* 10 November 2026 ; *Published:* 21 November 2026
https://doi.org/10.63328/IJRDES-V8RI6P9
https://www.youtube.com/watch?v=4Og5ArvMAtU
https://archive.org/details/jsp-ijrdes-v8ri6p9
Original Paper : Classification Research – Experimental Research Paper

Nearest Neighbours and Noise Features: An Exact Accuracy Formula and the Exponential Cost of Irrelevant Dimensions

Praveen Kumar Maharana

Department of Renewable Energy Systems , Nordhausen university of Applied sciences, Nordhausen, Germany.

* *Corresponding Author : Praveen Kumar Maharana ; praveenmaharana78933@gmail.com*

Abstract: Nearest-neighbour classifiers degrade when irrelevant features are added, and the usual explanation is that distances concentrate in high dimensions. We make this exact for the one-nearest-neighbour rule in a Gaussian model with s signal dimensions, p noise dimensions and class separation d. Conditioning on the test point, the squared distances to training points are non-central chi-square variables, and the accuracy is a two-dimensional integral over the test point's noise norm and signal offset (Eq. 4). The formula agreed with simulation to within 1.9 points in all 72 cells of a grid with 0 to 1,000 noise features and 20 to 5,000 training cases (mean absolute difference 0.7 points). It shows that the number of training cases needed for a fixed accuracy grows exponentially with the number of noise features, approximately as $\exp(p/d^2)$: for $d = 3$, 80% accuracy needs 14 cases with 10 noise features, 61 with 20, 2,898 with 50 and 1.7 million with 100. With 1,000 noise features the nearest neighbour has the right class for only about 61% above chance even with 5,000 training cases. On the digits dataset, adding 1,000 noise features reduced 1-NN accuracy from 97% to 40% with 1,300 training cases, while logistic regression and a random forest fell far less. The formula, heat maps and contour plots in this paper replace the qualitative curse-of-dimensionality argument by numbers. Synthetic results are simulations; the real data are one small benchmark set with appended noise.

Keywords: nearest neighbour, curse of dimensionality, irrelevant features, non-central chi-square, sample complexity

1. Introduction

The nearest-neighbour classifier is the oldest nonparametric rule [1], and it is consistent: with enough training data its error approaches at most twice the Bayes error in any dimension [1, 5, 6]. It also has a reputation for failing in high dimensions. The reputation rests on a geometric argument, that distances between points concentrate so that the nearest and the farthest neighbour become similar [2, 3, 7], and on the older warning of Bellman about the curse of dimensionality [4]. The argument is qualitative, and it is often misapplied. Distance concentration affects data whose coordinates are all informative in the same way; for a classifier the relevant question is whether the nearest neighbour has the right class, and that depends on how much of the distance comes from the features that carry the class and how much from the rest. Practitioners usually meet the problem as many irrelevant columns: an identifier-like feature, a block of unrelated measurements, a one-hot encoding of a high-cardinality variable. How many such columns can a nearest-neighbour method survive, and how much data does each one cost?

We answer this exactly in a Gaussian model. Given the test point, every training point is an independent draw, so the accuracy of the one-nearest-neighbour rule is a probability about the minimum of independent non-central chi-square variables. This gives a formula that agrees with simulation without any asymptotic approximation, and from it a rule for the data needed as a function of the number of noise features. We make five contributions.

- An exact expression for the accuracy of the 1-NN rule with signal and noise dimensions, evaluated by a two-dimensional integral (Section 4).
- A verification against simulation on a grid of 72 settings, and a map of the accuracy as a function of training size and noise dimensions.
- A measurement of the exponential growth of the required training size with the number of noise

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P9

features, with an approximate rule $\exp(p/d^2)$.

- A description of what goes wrong: the distributions of the nearest same-class and other-class distances, and the composition of the neighborhood of a test point.
- A real-data check on the digits dataset with appended noise features, comparing 1-NN with 5-NN, logistic regression and a random forest.

All synthetic experiments are simulations from a known Gaussian model. The real-data experiment appends independent noise columns to a standard benchmark.

2. Background and Related Work

Nearest neighbours. Cover and Hart showed that the asymptotic error of the 1-NN rule is at most twice the Bayes error [1]; consistency of k-NN and its rates are treated in standard texts [5, 6, 16], and Pestov questions the claim that k-NN must suffer from high dimension in general [9]. Kpotufe showed that k-NN regression adapts to the local intrinsic dimension [10], which explains why it works on data with low-dimensional structure, and Levina and Bickel give an estimator of that dimension [12].

Distance concentration. Beyer and colleagues asked when "nearest neighbour" is meaningful and gave conditions under which the ratio of farthest to nearest distance tends to one [2]. Aggarwal and colleagues compared distance metrics in high dimension [3], Hinneburg and colleagues studied the nearest neighbour in high-dimensional spaces [7], and Radovanović and colleagues described hubness [8]. Friedman proposed to adapt the metric to the class structure [13], and Hastie and colleagues discuss the curse of dimensionality [11].

What we add. These works establish when neighbours lose meaning in general. We give the exact accuracy of the 1-NN rule when a few features carry the class and the rest are noise, which is the situation of feature-rich tabular data, and we express the cost of noise features in training cases.

3. Research Methodology and Gap

3.1. The Gaussian Model

A case is a vector in $\mathbb{R}^{s+p}$. The first s coordinates are signal and the remaining p are noise. The class $y \in \{0, 1\}$ is balanced. All coordinates are unit-variance Gaussian, independent given the class; the two class means differ by d in total, along the first coordinate, and are equal elsewhere. The Bayes accuracy of this problem is $\Phi(d/2)$, which is 93% for $d = 3$ whatever p. A 1-NN classifier has n training points, $n/2$ from each class, and uses Euclidean distance. We take $s = 2$ and $d = 3$ unless stated.

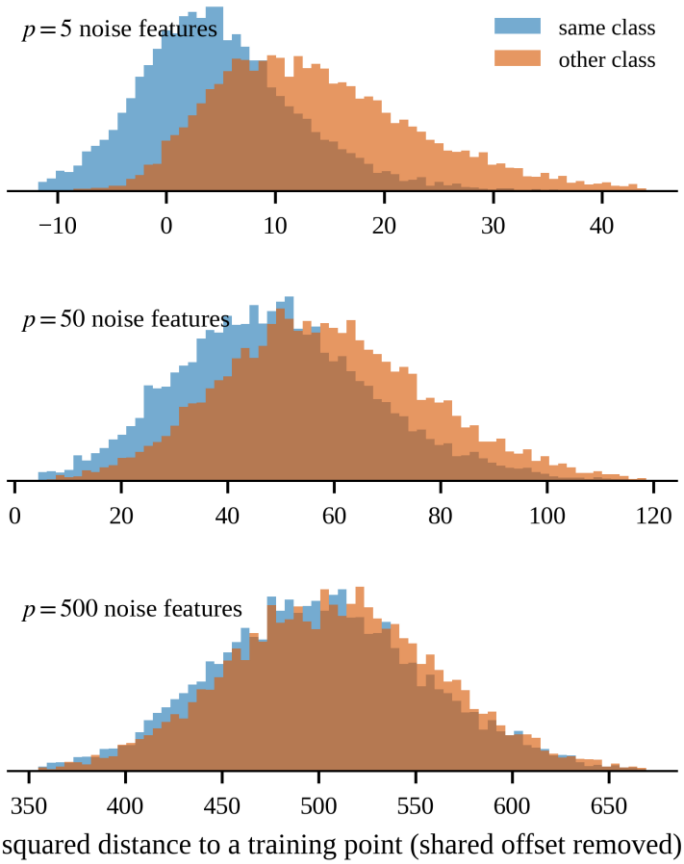


Figure. 1 Squared distances from a test point to training points of its own class and of the other class ($d = 3$, two signal dimensions). As noise features are added, the two distributions merge. The shared term $\|x\|^2$ is removed.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P9

3.2. Real Data and Other Classifiers

The real-data study uses the digits dataset (1,797 cases, ten classes, 64 features standardised) [15] from scikit-learn [14], with p independent $N(0, 1)$ columns appended. We compare 1-NN, 5-NN, logistic regression ($C = 0.1$) and a random forest of 50 trees.

4. Proposed System

Proposition 1 (exact 1-NN accuracy). *Condition on a test point x of class 0 with class mean μ. Write $x - \mu = (\xi, z)$, with $\xi \sim N(0, 1)$ the coordinate along the class-separation direction and $R^2 = \|z\|^2 \sim \chi^2_{s+p-1}$ the rest. Let D_0 be the squared distance from x to a class-0 training point, and D_1 to a class-1 point. Given x,*

$$D_0 \sim \chi'^2_{s+p}(\lambda_0), \quad D_1 \sim \chi'^2_{s+p}(\lambda_1), \quad \lambda_0 = \xi^2 + R^2, \quad \lambda_1 = (\xi + d)^2 + R^2, \tag{1}$$

where χ'^2 is the non-central chi-square distribution, and the training points are independent. The 1-NN rule is correct when the smallest distance belongs to a class-0 point. Hence, with $n_0 = n_1 = n/2$,

$$\mathrm{Acc} = E_{\xi, R^2}\left[\int_0^1 n_0(1-u)^{n_0-1}\, S_1\left(F_0^{-1}(u)\right)^{n_1} du \right], \tag{2}$$

where F_0 is the distribution function of D_0 and $S_1 = 1 - F_1$ the survival function of D_1 (both given ξ and R^2).

Proof. Given x, the n distances are independent, so the probability that the minimum over class 0 is below every class-1 distance is the integral of the density of the class-0 minimum, which is $n_0(1 - u)^{n_0-1}$ in the quantile scale $u = F_0(t)$, times the probability that all n_1 class-1 distances exceed it.

The squared distance from x to a point with mean μ' and unit covariance is non-central chi-square with noncentrality $\|x - \mu'\|^2$. The term that is common to all distances, $\|x\|^2$, cancels in the comparison and is why only R^2 and ξ matter.

The two random quantities ξ and R^2 are independent, so the outer expectation is a two-dimensional integral which we evaluate by Gauss–Hermite quadrature in ξ and by quantiles of R^2. Both distance distributions have the same degrees of freedom $s + p$; the classes differ only in the noncentrality, by $d^2 + 2d\xi$, so the signal enters as a random shift of size d^2 with a random part of standard deviation $2d$.

Remark (consequences). The conditional standard deviation of each distance is about $\sqrt{(6(s + p))}$ and grows with the number of noise features, while the signal shift d^2 does not.

The probability that a single other-class distance falls below a single same-class distance therefore tends to 1/2 as $p \to \infty$. With n points the minimum of each class shrinks to the extreme tail, and the minimum of the same-class group still falls below the minimum of the other-class group provided n is large enough; the required n is the subject of Section 6.3.

$$\mathrm{Acc}(p, n) \to 1/2 \ \text{ as } \ p \to \infty \text{ at fixed } n; \quad \mathrm{Acc}(p, n) \to \mathrm{Acc}_{1\text{-NN}}(d) \ \text{ as } \ n \to \infty \text{ at fixed } p. \tag{3}$$

Eq. (3) states the two limits. The first is the failure of the method at fixed sample size, and the second is its consistency [1]. The interesting question is the size of n that moves between them.

Rule of thumb (empirical). *For the accuracy target a, the training size $n^*(p)$ satisfying $\mathrm{Acc}(p, n^*) = a$ behaves like*

$$\log n^* \approx \kappa(d)\, p, \quad \kappa(d) \approx 1/d^2. \tag{4}$$

We found Eq. (4) by evaluating the exact formula, not by derivation; the fitted values of κ for $d = 2, 3, 4$ are given in Section 6.3.

5. Experimental Setup and Implementation

The simulation grid has $p \in \{0, 5, 10, 20, 50, 100, 200, 500, 1{,}000\}$ and $n \in \{20, 50, 100, 200, 500, 1{,}000, 2{,}000, 5{,}000\}$, with $s = 2$ and $d = 3$. Each cell uses 12 repetitions (6 for the largest cells) of a fresh training set and 300 fresh test points. The exact formula is evaluated for the same grid. The sample-size study solves $\mathrm{Acc}(p, n) = a$ for n with a root finder, for $d = 2$ with $a = 0.7$, $d = 3$ with $a = 0.8$ and $d = 4$ with $a = 0.85$, and checks two solutions by simulation (20 repetitions). The distance distributions use 40 test points against 2,000 training points. The distance-ratio study uses $n = 500$, 900 test points per p. The neighbourhood study uses $n = 500$, $k = 5$ and six repetitions per p.

The real-data study uses 497 random test cases and, for each of four repetitions, training sets of $n \in \{50, 100, 200, 400, 800, 1{,}300\}$ cases and $p \in \{0, 10, 30, 100, 300, 1{,}000\}$ appended noise features. The other classifiers are trained at $n = 200$ and 800.

6. Results and Analysis

6.1. The Formula Against Simulation

Figures 2 and 3 compare the simulated accuracy with Eq. (2). The two agree in every one of the 72 cells, with maximum absolute difference 1.9 points and mean absolute difference 0.7 points, which is of the order of the simulation noise (the standard deviation of an average of 12 repetitions of 300 test points is about 1 point).

The accuracy is 90% without noise features and 5,000 training cases (the formula gives 90%; the Bayes accuracy is 93%), 85% with 20 noise features, 75% with 100 and 61% with 1,000, where chance is 50%. At $n = 100$ the numbers are 90%, 83%, 71% and 58%. Adding data helps less and less as noise features are added: at $p = 1{,}000$ going from 20 to 5,000 training cases raises accuracy by only 6 points, and at $p = 0$ by 0.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P9

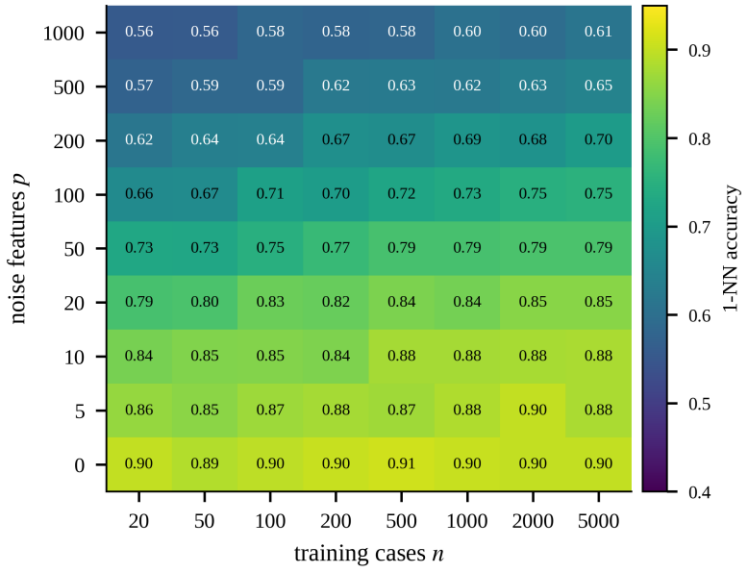


Figure 2: Simulated 1-NN accuracy for $d = 3$ as a function of the number of training cases and noise features (12 repetitions per cell, 300 test points).

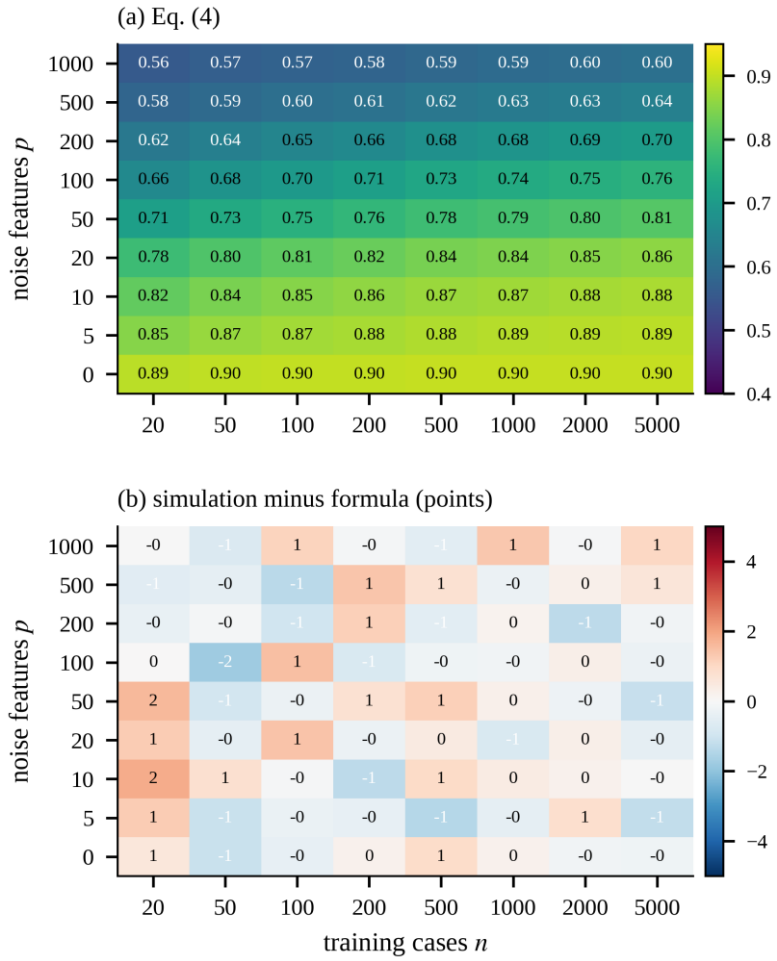


Figure 3: (a) Exact 1-NN accuracy from Eq. (4). (b) Simulation minus formula, in percentage points.

Table 1: 1-NN accuracy (%): simulation and Eq. (2) for selected cells ($d = 3$).

p	n = 100 Sim	n = 100 Eq. (2)	n = 1,000 Sim	n = 1,000 Eq. (2)	n = 5,000 Sim	n = 5,000 Eq. (2)
0	90	90	90	90	90	90
10	85	85	88	87	88	88
50	75	75	79	79	79	81
200	64	65	69	68	70	70
1000	58	57	60	59	61	60

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P9

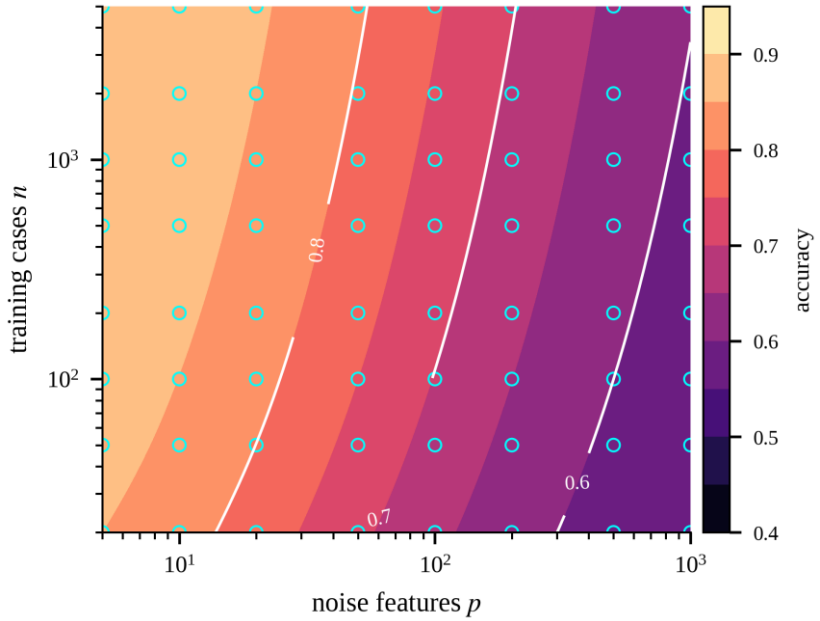


Figure. 4 Accuracy contours in the plane of noise features and training cases, from a smooth interpolation of Eq. (4); circles show the simulated cells.

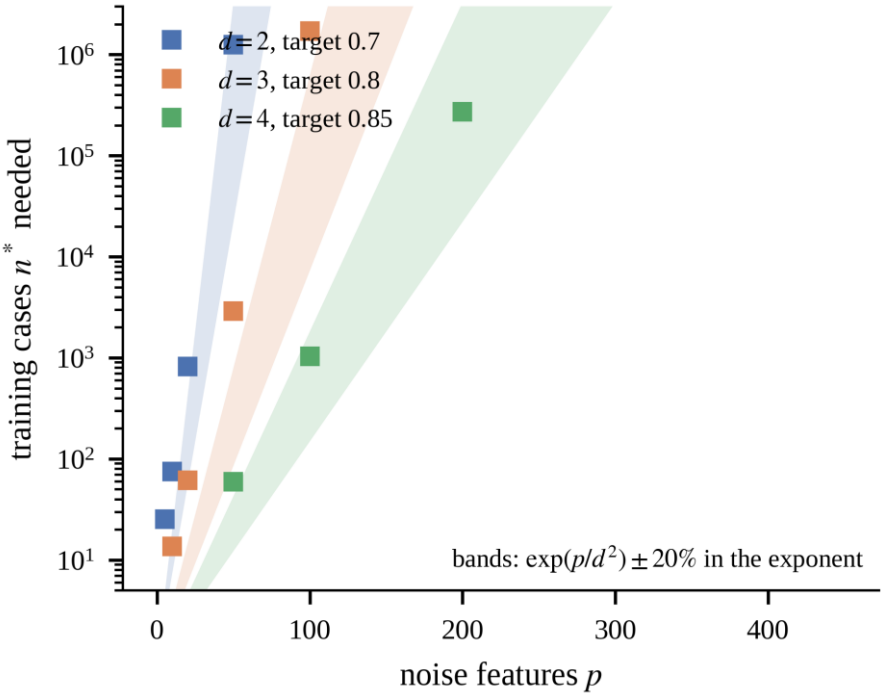


Figure. 5 Training cases needed for a target 1-NN accuracy against the number of noise features (squares: Eq. 4), for three class separations. The bands show $\exp(p/d^2)$ with the exponent varied by ±20%.

6.2. What Goes Wrong

Figure 1 shows the mechanism. With 5 noise features the distribution of squared distances to other-class points is shifted to the right of the same-class distribution by about $d^2 = 9$; with 500 the two distributions overlap almost completely, because their spread grows as $\sqrt{p}$ while the shift stays at d^2. The nearest neighbour, however, depends on the lower tails. Figure 6 shows the ratio of the nearest same-class to the nearest other-class distance for $n = 500$. Its median is 0.16 without noise features, 0.79 with 10, 0.95 with 50, and 0.99 and 1.00 with 200 and 1,000; at 1,000 the ratio is distributed symmetrically around 1 and the nearest neighbour is a coin flip. In terms of neighbourhoods (Figure 7), the share of test points whose five nearest neighbours all belong to their own class falls from 76% with no noise features to 61% with 10, 36% with 50 and 6% with 1,000.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P9

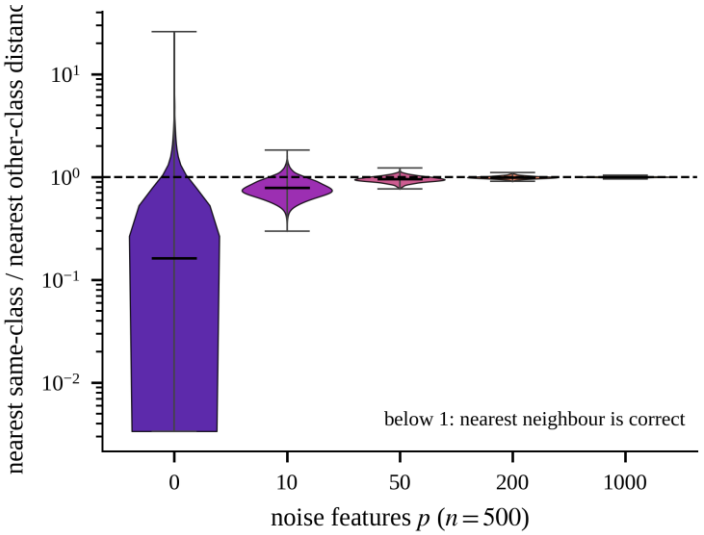


Figure. 6 Distribution over test points of the ratio of the nearest same-class distance to the nearest other-class distance (n = 500, d = 3). A ratio below 1 means the nearest neighbour has the right class.

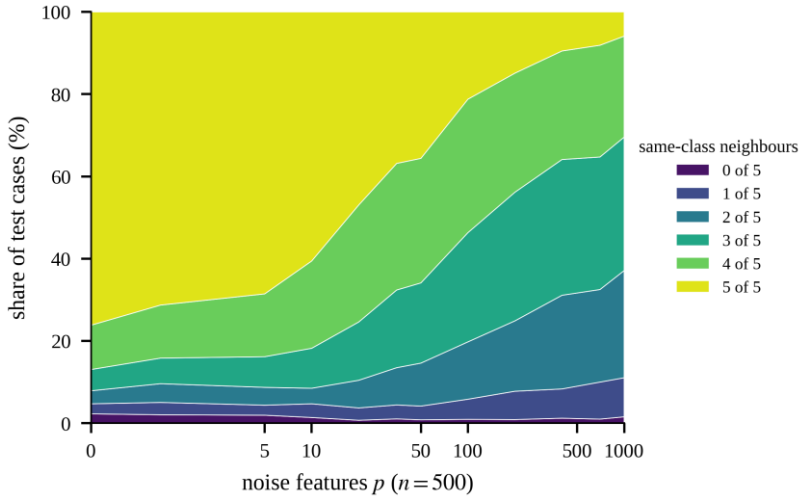


Figure. 7 Among the five nearest neighbours of a test point, the share of test cases with 0 to 5 neighbours of its own class, as noise features are added (n = 500, d = 3).

6.3. How Much Data Does Each Noise Feature Cost?

Figure 5 and Table 2 give the number of training cases needed to reach a target accuracy. For d = 3 and a target of 80%, 1-NN needs 14 cases with 10 noise features, 61 with 20, 2,898 with 50 and 1.7 million with 100. Each additional ten noise features multiplies the requirement by about 3.7. The simulation check confirms the formula: at the computed sizes the simulated accuracy is 81% (p = 10, n = 14) and 80% (p = 50, n = 2,898) against a target of 80%. The growth is exponential: the fitted slope of log n^* on p is 0.241 for d = 2, 0.130 for d = 3 and 0.056 for d = 4, against $1/d^2$ = 0.250, 0.111 and 0.063. The rule $\exp(p/d^2)$ is a rough guide, within a factor of about 1.2 on the exponent. The practical meaning is that a nearest-neighbour method can tolerate a handful of noise features per signal dimension and no more.

Table. 2 Training cases needed for a target 1-NN accuracy (Eq. 2): d = 2 for 70%, d = 3 for 80%, d = 4 for 85%; – means the target is not reached below two million cases.

p	d = 2 (70%)	d = 3 (80%)	d = 4 (85%)
5	25	–	–
10	75	14	–
20	820	61	–
50	1.25 million	2,898	59
100	–	1.7 million	1,033
200	–	–	0.27 million

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P9

6.4. Real Data

Figure 8 shows 1-NN accuracy on digits with appended noise features. With 1,300 training cases accuracy falls from 97% without noise features to 92% with 30, 83% with 100, 64% with 300 and 40% with 1,000; with 50 training cases the corresponding numbers are 76%, 66%, 55%, 39% and 23%. The contour pattern is the one in Figure 4: the lines of constant accuracy move to larger n as p increases, and for $p \geq 300$ no training size in the range reaches 70%. Figure 9 compares four classifiers with 800 training cases: 1-NN falls from 96% to 38%, 5-NN from 96% to 46%, logistic regression from 96% to 81% and the random forest from 97% to 88%. Averaging over five neighbours helps (8 points at $p = 1{,}000$), but does not change the trend, whereas the forest, which selects features, is much less affected.

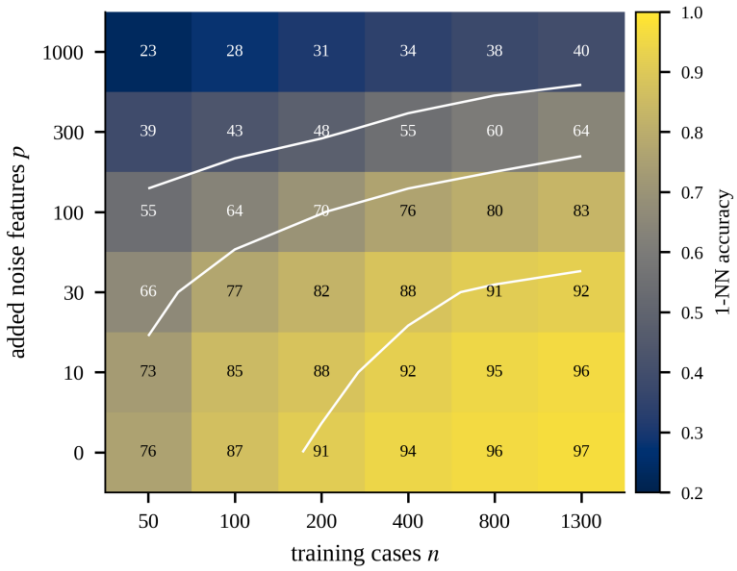


Figure. 8 Real digits (64 features) with added noise features: 1-NN test accuracy (%), mean of four random splits. White lines mark 50%, 70% and 90%.

Table. 3 Real digits, 800 training cases: test accuracy (%) with appended noise features.

Classifier	$p = 0$	$p = 10$	$p = 100$	$p = 1{,}000$
1-NN	96	95	80	38
5-NN	96	95	86	46
Logistic	96	96	93	81
Forest	97	96	95	88

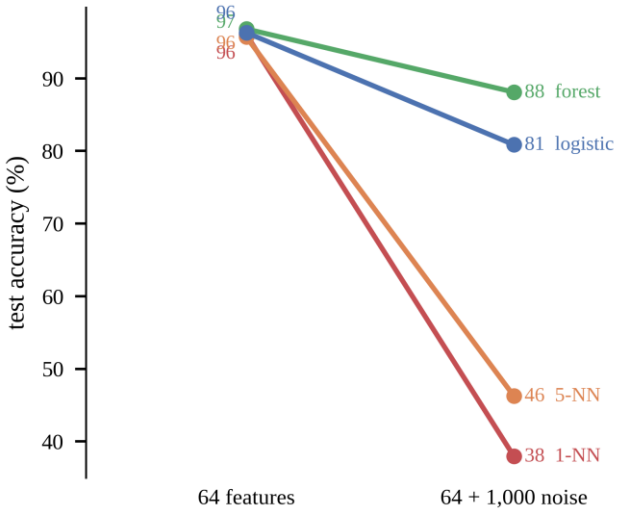


Figure. 9 Test accuracy on digits with 800 training cases, without noise features and with 1,000 added: four classifiers.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P9

6.5. A practical recipe

The results suggest a short procedure. First, count the features that could carry the class and the ones that could not; a nearest-neighbour method can tolerate roughly as many noise features as the squared class separation allows (Eq. 4). Second, if the count is larger, select or weight features before computing distances, as the random forest does implicitly; the benefit is exponential in the number removed. Third, use Table 2 or the formula to estimate the training size a given number of noise features would need, and treat a requirement above the available data as a reason to reduce the dimension, not to collect more. Fourth, check the ratio of the nearest same-class to the nearest other-class distance on held-out data: a median near 1 means the neighbours are uninformative. Fifth, averaging over more neighbours does not repair the problem.

Discussion and Limitations

What we think carries over. Three points seem general. A nearest-neighbour classifier does not degrade gradually: the data needed grows exponentially in the number of noise dimensions, so the same method can work with 10 noise features and fail with 50. The loss is a loss of ranking, not of distances: the ratio of nearest-neighbour distances shows it directly. And the cost is set by the signal-to-noise structure, measured by d^2, not by the raw dimension.

Synthetic setting. The model has Gaussian classes with equal covariance, independent noise coordinates with the same scale as the signal coordinates, and a signal confined to one direction. Correlated noise, noise with smaller variance, redundancy among features and class overlap in many directions will change the numbers, and low-dimensional structure in the noise would help [10]. The rule of Eq. (4) is empirical, fitted to three values of d. The formula is for 1-NN with balanced classes; k-NN requires the kth order statistic.

Real data. We appended independent Gaussian columns to one dataset. Real irrelevant features are not independent of one another and are not Gaussian, and they may be correlated with the signal. We used standardised features, so the added noise has the same scale as the original features; unscaled data would give the noise a different weight. We did not study metric learning or feature weighting methods [13].

7. Conclusion

For a Gaussian model with few signal dimensions and many noise dimensions, the 1-NN accuracy is an exact two-dimensional integral over non-central chi-square distances, which matched simulation within 1.9 points over 72 settings. The number of training cases needed for a given accuracy grows exponentially with the number of noise features, approximately as $\exp(p/d^2)$: for $d = 3$ it rises from 14 cases at 10 noise features to 1.7 million at 100. On digits, appending 1,000 noise features cut 1-NN accuracy from 97% to 40% while a random forest fell to 88% at 800 cases. The next steps are to extend the formula to k neighbours and correlated noise, to derive the rate $\kappa(d)$, and to study feature-weighted metrics.

References

[1]. T. Cover, P. Hart. Nearest neighbour pattern classification. *IEEE Trans. Information Theory*, 13(1):21–27, 1967.

[2]. K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft. When is "nearest neighbour" meaningful? In *ICDT*, 1999.

[3]. C. C. Aggarwal, A. Hinneburg, D. A. Keim. On the surprising behaviour of distance metrics in high dimensional space. In *ICDT*, 2001.

[4]. R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[5]. L. Devroye, L. Györfi, G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[6]. C. J. Stone. Consistent nonparametric regression. *Ann. Statist.*, 5(4):595–620, 1977.

[7]. A. Hinneburg, C. C. Aggarwal, D. A. Keim. What is the nearest neighbour in high dimensional spaces? In *VLDB*, 2000.

[8]. M. Radovanović, A. Nanopoulos, M. Ivanović. Hubs in space: popular nearest neighbours in high-dimensional data. *J. Mach. Learn. Res.*, 11:2487–2531, 2010.

[9]. V. Pestov. Is the k-NN classifier in high dimensions affected by the curse of dimensionality? *Computers & Mathematics with Applications*, 65(10):1427–1437, 2013.

[10]. S. Kpotufe. k-NN regression adapts to local intrinsic dimension. In *NIPS*, 2011.

[11]. T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning*, 2nd ed. Springer, 2009.

[12]. E. Levina, P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *NIPS*, 2004.

[13]. J. H. Friedman. Flexible metric nearest neighbour classification. Technical report, Stanford University, 1994.

[14]. F. Pedregosa et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[15]. E. Alpaydin, C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998.

[16]. G. Biau, L. Devroye. *Lectures on the Nearest Neighbor Method*. Springer, 2015.

A. Derivations

Eq. (1). For $y \sim N(\mu', I_m)$ with $m = s + p$, the vector $x - y$ has mean $x - \mu'$ and covariance I. Its squared norm is non-central chi-square with m degrees of freedom and noncentrality $\|x - \mu'\|^2$. For a class-0 point $\mu' = \mu$, giving $\lambda_0 = \|x - \mu\|^2 = \xi^2 + R^2$. For a class-1 point, $\|x - \mu'\|^2 = \|x - \mu\|^2 + d^2 + 2d\xi = (\xi + d)^2 + R^2$, where ξ is the component of $x - \mu$ along $\mu' - \mu$.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P9

Eq. (2). For independent class-0 distances with distribution function F_0, the probability density of their minimum at t is $n_0\, f_0(t)(1 - F_0(t))^{n_0-1}$. Changing variable to $u = F_0(t)$, and multiplying by the probability $S_1(t)^{n_1}$ that all class-1 distances are larger, gives the integral in Eq. (2). The outer expectation integrates over $\xi \sim N(0, 1)$ and $R^2 \sim \chi^2_{m-1}$.

Limits (Eq. 3). At fixed n, as $p \to \infty$ the standard deviation of each distance grows as $\sqrt{p}$ while the mean difference stays at d^2, so each class-1 distance is below each class-0 distance with probability tending to 1/2, and the minimum is equally likely to come from either class. At fixed p, as $n \to \infty$ the 1-NN accuracy converges to its asymptotic value by Cover and Hart [1].

B. Parameters

Synthetic: $s = 2$, $d = 3$ (grid), $p \in \{0, 5, 10, 20, 50, 100, 200, 500, 1{,}000\}$, $n \in \{20, 50, 100, 200, 500, 1{,}000, 2{,}000, 5{,}000\}$; 12 repetitions (6 for cells with $n(p + 2) \geq 2 \times 10^6$), 300 test points; formula by 15-node Gauss–Hermite quadrature in ξ, 12 quantile nodes for R^2, and a 501-point quantile grid for u. Sample-size study: root finding in log n to tolerance 0.02. Real data: digits standardised, 497 test cases, four repetitions, $n \in \{50, 100, 200, 400, 800, 1{,}300\}$, $p \in \{0, 10, 30, 100, 300, 1{,}000\}$; 5-NN, logistic regression ($C = 0.1$, 300 iterations) and a random forest with 50 trees at $n = 200$ and 800. All figures were produced by code at 600 dpi and are printed at their stored width. Code, results and figures are in the supplementary archive.

How To Cite:

Praveen Kumar Maharana , " Nearest Neighbours and Noise Features: An Exact Accuracy Formula and the Exponential Cost of Irrelevant Dimensions " , International Journal of Research and Development in Engineering Sciences (IJRDES), Volume 8 , Issue 6 , pp 88 – 96, November 2026.
CrossRef DOI: https://doi.org/10.63328/ijrdes-v8ri6p9

Declaration

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

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Generative AI Statement: The authors declare that generative AI was not used in the preparation or creation of this manuscript. The alternative text (alt text) accompanying the figures in this article was generated by Frontiers with the assistance of artificial intelligence. Reasonable efforts have been made to ensure the accuracy and appropriateness of the generated alt text, including review by the authors wherever possible. If you identify any inaccuracies or issues, please contact the editorial team.

Publisher's Note: The statements, opinions, and claims presented in this article are exclusively those of the authors and do not necessarily reflect the views of their affiliated institutions, the publisher, editors, or reviewers. Any products discussed or evaluated, as well as any claims made by their manufacturers, are not warranted, approved, or endorsed by the publisher.

Acknowledgements

We thank everyone who inspired our work

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com