Journal of Innovation in Applied Natural Science
Volume 2, Issue 2, 80-89.
e_ISSN: 3110-4150
https://e-journal.gomit.id/jinas
DOI: 10.58723/jinas.v2i2.223

Comparative Analysis of K-Nearest Neighbor and Fuzzy K-Nearest Neighbor for Diabetes Classification Using the Pima Indians Diabetes Dataset

Ulfasari Rafflesia*
Universitas Bengkulu, INDONESIA

Siska Dwi Kumala
Universitas Bengkulu, INDONESIA

Ratna Widayati
Universitas Bengkulu, INDONESIA

Aisyah Nooravieta S.
Universitas Bengkulu, INDONESIA

Oon Septa
Universitas Bengkulu, INDONESIA

*** Corresponding author:**
Ulfasari Rafflesia, Universitas Bengkulu, INDONESIA. ✉ ulfasari@unib.ac.id

Article Info

Article history:

Received: August 4, 2026
Revised: August 19, 2026
Accepted: September 8, 2026

Keywords:

Classification
Diabetes prediction
Fuzzy KNN
K-Nearest Neighbor
Machine learning

Abstract

Background of study: Diabetes mellitus remains one of the most common chronic diseases worldwide; thus, early and accurate prediction is critical for appropriate intervention.
Aims: This study examines the performance of the standard K-Nearest Neighbor (KNN) method and its fuzzy counterpart, Fuzzy K-Nearest Neighbor (Fuzzy KNN), in classifying diabetes status and compares their behavior across different neighborhood sizes (k).
Methods: The Pima Indians Diabetes Dataset (768 cases, 8 clinical variables) was used. Zero values in physiologically implausible attributes (glucose, blood pressure, skin thickness, insulin, and BMI) were treated as missing values based on the physiological interpretation of these variables and imputed using class-wise medians. The dataset was then split into 80% training and 20% testing subsets using stratified sampling to preserve class proportions. Five-fold cross-validation on the training set was used to determine the optimal number of neighbors (k), after which both models were evaluated on the held-out test set using accuracy, precision, recall, and F1-score.
Result: The optimal k varied between models: k=15 for KNN and k=3 for Fuzzy KNN. Cross-validation accuracy across the tested k range (3–21) remained relatively stable for both methods (approximately 0.80–0.83), indicating that performance was not highly sensitive to k selection. On the test set, Fuzzy KNN outperformed conventional KNN across all evaluation metrics (accuracy: 0.779 vs. 0.773; precision: 0.679 vs. 0.673; recall: 0.704 vs. 0.685; F1-score: 0.691 vs. 0.679).
Conclusion: For the Pima Indians Diabetes Dataset, Fuzzy KNN achieved slightly better performance than conventional KNN across all evaluated metrics on the held-out test set. These findings indicate that distance-based membership weighting may provide a modest advantage over crisp majority voting for diabetes classification in this dataset.

To cite this article: Rafflesia, U, Kumala, S. D., Widayati, R., Setiawan, A. N., and Septa, O. (2026). Comparative Analysis of K-Nearest Neighbor and Fuzzy K-Nearest Neighbor for Diabetes Classification Using the Pima Indians Diabetes Dataset. *Journal of Innovation in Applied Natural Science, 2*(2),80-89

INTRODUCTION

Diabetes mellitus continues to be one of the most prevalent chronic diseases worldwide, and early, accurate detection remains essential for timely clinical intervention (Addou et al., 2026; Pikula et al., 2024). Machine learning classifiers have become widely used tools for medical diagnosis support, with the K-Nearest Neighbor (KNN) algorithm being particularly attractive due to its conceptual simplicity, ease of implementation, and lack of a training phase (Sachdeva et al., 2025; Stefane Souza & Araújo Lima, 2024). However, standard KNN classifies instances through simple majority voting, which can perform poorly when class boundaries overlap a common characteristic of real-world clinical datasets such as diabetes records, where diagnostic features often show considerable overlap between diabetic and non-diabetic cases. To address this limitation, Fuzzy K-Nearest Neighbor (Fuzzy KNN) incorporates distance-based membership weighting rather than relying solely on crisp majority voting. This raises the question of whether Fuzzy KNN provides a consistent performance advantage over standard KNN and whether the optimal neighborhood size (k) differs between the two methods.

KNN has been extensively studied and modified to improve its performance in classification tasks (Amer et al., 2025; Zhang, 2022). Recent reviews have identified several persistent challenges in KNN, including sensitivity to noise, the curse of dimensionality, and the selection of an appropriate neighborhood size (k) (Halder et al., 2024; Suyal & Goyal, 2022; Syriopoulos et al., 2025). These challenges are particularly relevant to medical classification, where observations may exhibit overlapping feature distributions and heterogeneous local patterns. Accordingly, KNN and its variants have been applied to various diagnostic problems, including Alzheimer's disease, breast cancer, and lung cancer (Dinesh et al., 2024; Lahmiri, 2023; Sachdeva et al., 2025). Collectively, these studies demonstrate the continued applicability of neighborhood-based classification in medical diagnosis while highlighting the importance of how local neighborhood information is represented and weighted.

One approach to addressing the limitations of crisp neighborhood voting is Fuzzy KNN, originally introduced by (González-Martínez et al., 2020), which assigns class membership degrees to observations based on their neighboring instances and distances rather than relying solely on majority voting. Subsequent research has extended the original formulation through generalized distance functions, adaptive neighbor selection, and alternative aggregation mechanisms (Bian et al., 2022; Mailagaha Kumbure et al., 2020; Mailagaha Kumbure & Luukka, 2022). These developments indicate a broader trend toward incorporating uncertainty and distance information into neighborhood-based classification. In medical applications, Fuzzy KNN and related fuzzy approaches have been investigated for heart disease, Parkinson's disease, and diabetes diagnosis (Chandrasekaran et al., 2023; Kaur & Khehra, 2022; Salem et al., 2022). In particular, (Salem et al., 2022) applied a fine-tuned Fuzzy KNN based on uncertainty membership to the Pima Indians Diabetes Dataset, illustrating how the original fuzzy framework can be further enhanced for diabetes classification.

Existing studies have established the relevance of both KNN and Fuzzy KNN for diabetes classification using the Pima Indians Diabetes Dataset. KNN has been evaluated as one of several conventional supervised classifiers (Choubey et al., 2020;Tasin et al., 2023), while Fuzzy KNN has been applied using feature-selection and fuzzy modeling strategies (Li et al., 2023; Vommi & Krishna, 2024). Other studies have incorporated both KNN and Fuzzy KNN within hybrid or optimized frameworks. For example, (Arowolo et al., 2021)combined KNN and Cuckoo-Fuzzy-KNN with Firefly and Cuckoo Search-based feature selection, while (Salem et al., 2022) directly evaluated KNN and Fuzzy KNN alongside a tuned Fuzzy KNN based on uncertainty membership. Investigated Fuzzy KNN within a modified artificial immune recognition framework (Azevedo et al., 2024; Mayawi & Subanar, 2025). Representative studies on KNN and Fuzzy KNN for diabetes classification are summarized in Table 1, highlighting their datasets, methodological characteristics, and relevance to the present study.

Table 1. State of the art comparison of KNN and Fuzzy KNN for diabetes classification

Study	Dataset	Method	Main methodological characteristic	Relevance to the present study
(González-Martínez et al., 2020)	Pima Indians Diabetes	Fuzzy KNN	Fuzzy modeling combined with feature selection	Shows an application of Fuzzy KNN to the Pima dataset, but focuses on feature selection and fuzzy modeling

(Azevedo et al., 2024; Mayawi & Subanar, 2025)	Pima Indians Diabetes	AIRS2 with Fuzzy KNN	Fuzzy KNN incorporated into an artificial immune recognition framework	Uses Fuzzy KNN in a hybrid framework rather than comparing the original KNN and Fuzzy KNN
(Lohrmann et al., 2025)	Pima Indians Diabetes	KNN and Cuckoo-Fuzzy-KNN	Search-based feature selection combined with KNN-based methods	Includes both KNN and a Fuzzy KNN variant, but within feature-selection strategies
(Arowolo et al., 2021)	Pima Indians Diabetes	KNN and other classifiers	Comparative evaluation of conventional supervised classifiers	Demonstrates the use of KNN for diabetes classification, but does not include Fuzzy KNN
(Salem et al., 2022)	Pima Indians Diabetes	KNN, Fuzzy KNN, and tuned Fuzzy KNN	Uncertainty-based membership and grid-search optimization	Provides a direct KNN–Fuzzy KNN comparison, but focuses on an enhanced/tuned Fuzzy KNN
Present study	Pima Indians Diabetes	KNN and original Fuzzy KNN	Same preprocessing and evaluation framework; systematic comparison across k = 3–21	Focuses on the baseline difference between crisp majority voting and distance-based fuzzy membership without additional algorithmic enhancement

As summarized in Table 1 the existing studies generally evaluate KNN and Fuzzy KNN within additional feature-selection, hybridization, uncertainty-based membership, or optimization procedures. Consequently, the isolated difference between standard crisp majority voting and the original distance-based membership mechanism of Fuzzy KNN is not the primary focus of these studies. Moreover, the effect of neighborhood size (k) is typically considered as part of model tuning rather than as the main object of a systematic comparison between the two original formulations. Therefore, a controlled baseline comparison of standard KNN and the original (Hameed & Joshi, 2024; Lohani et al., 2025)Fuzzy KNN under a common experimental framework provides a useful complementary perspective for clarifying how the two voting mechanisms behave across the same range of k values. The present study addresses this gap by comparing the two methods on the Pima Indians Diabetes Dataset using the same preprocessing and evaluation framework, with k systematically examined from 3 to 21. This rigorous benchmarking incorporates stratified cross-validation to ensure consistent performance evaluation across different neighborhood sizes (Aria et al., 2025).

A straightforward and reproducible comparison of KNN and Fuzzy KNN is valuable both pedagogically and practically because it isolates the effect of distance-weighted membership under standard, non-optimized conditions. Such a comparison can clarify whether the use of distance-based membership provides an advantage over conventional majority voting and offers practical insight into neighborhood-size selection. This baseline comparison is particularly relevant for researchers and practitioners seeking transparent and interpretable models before adopting more complex Fuzzy KNN variants.

This study aims to compare the classification performance of standard KNN and Fuzzy KNN in predicting diabetes status using the Pima Indians Diabetes Dataset, and to examine how the optimal number of neighbors (k) differs between the two methods. It is hypothesized that (H1) Fuzzy KNN will outperform standard KNN across accuracy, precision, recall, and F1-score due to its distance-based membership weighting, which better accommodates overlapping class boundaries; and (H2) the optimal k for Fuzzy KNN will be smaller than that for standard KNN, since its inherent distance weighting already mitigates the effect of noisy neighbors that standard KNN requires a larger k to smooth out.

METHOD

This study employed a quantitative, comparative computational design based on secondary data analysis. Rather than direct human-subject recruitment, the study compared the classification performance of two supervised machine learning algorithms standard K-Nearest Neighbor (KNN) and Fuzzy K-Nearest Neighbor (Fuzzy KNN) using a publicly available, de-identified clinical dataset.

No human participants were directly recruited for this study. The analysis used a publicly available, de-identified secondary dataset (the Pima Indians Diabetes Dataset), originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The dataset consists of clinical records from 768 female patients of Pima Indian heritage, aged 21 years or older. As the data were fully de-identified and publicly accessible prior to this study, no additional ethical approval or informed consent procedure was required for this secondary analysis.

The population represented in the dataset is limited to female patients of Pima Indian heritage aged ≥21 years, originally sampled by NIDDK. No additional sampling was performed by the authors; all 768 available records were used in full. The dataset does not involve a psychometric survey instrument; rather, it consists of eight clinical/laboratory measurements per patient (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age) and a binary diagnostic outcome (diabetic/non-diabetic), as established in the original data collection by (Baghdadi et al., 2023; Bonmati et al., 2021; Calster et al., 2023). As such, traditional validity/reliability coefficients (e.g., Cronbach's alpha) do not apply; the clinical validity of the measurements is inherited from the dataset's original medical data collection protocol.

Data analysis was carried out using Python 3, with the scikit-learn library for standard KNN implementation, preprocessing, cross-validation, and evaluation metrics. Fuzzy KNN was implemented manually following the original formulation of (Fan, 2025; Szczuko et al., 2021). For reproducibility, both classifiers were evaluated using the same preprocessing procedure, train–test split, Euclidean distance metric, candidate neighborhood sizes (k = 3–21), and five-fold cross-validation protocol. The Fuzzy KNN implementation followed the original formulation of (Akhtar et al., 2024; Siyar et al., 2020), with the fuzzifier parameter fixed at m = 2.

K-Nearest Neighbor (KNN). For a test instance x, the Euclidean distance to every training instance is computed as:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{p} (x_j - x_{ij})^2} \qquad (1)$$

where p is the number of features. The k training instances with the smallest distance to x are selected as its nearest neighbors, and x is assigned to the class that occurs most frequently among those k neighbors (majority voting):

$$y^{\wedge} = argmax \quad_c \sum_{i \in N} \mathbb{1}(y_i = c) \qquad (2)$$

Fuzzy K-Nearest Neighbor (Fuzzy KNN). Rather than a hard vote, Fuzzy KNN assigns each class a membership degree based on the inverse distance of x to its k nearest neighbors, following (Fadaei-Kermani & Ghaeini-Hessaroeyeh, 2020) :

$$\mu_c(x) = \frac{\sum_{i \in N} \mu_{ci} \cdot \|x - x_i\| \quad^{-2/(m-1)}}{\sum_{i \in N} \|x - x_i\| \quad^{-2/(m-1)}} \qquad (3)$$

where μci is the membership of training instance xi to class c, and m is the fuzzifier parameter controlling how strongly distance influences the weighting, m = 2, in this study, consistent with common practice (Bian et al., 2022; Mailagaha Kumbure & Luukka, 2022). The final predicted class is the one with the highest resulting membership value:

$$y^{\wedge} = argmax \quad_c \mu_c(x) \qquad (4)$$

In this study, the initial class membership μci of each training instance was determined using the crisp-neighbor method proposed by (W. Wu et al., 2021), in which an instance belonging to class c receives a higher membership toward c and a proportionally lower membership toward other classes, based on the class distribution among its own k nearest neighbors in the training set (Suárez et al., 2024).

Zero values in five physiologically implausible attributes (Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI) were treated as missing values and imputed using class-wise medians.

The imputation was performed separately within each outcome class to preserve the class-specific distribution of the variables. The preprocessed dataset was then divided into training (80%) and testing (20%) subsets using stratified sampling to preserve the original class proportions. Five-fold cross-validation was performed on the training set to identify the optimal number of neighbors (k) for each model, tested across k = 3 to 21. The held-out test set was not used for model selection or determination of the optimal k and was reserved exclusively for the final evaluation of each model. Each model was then retrained on the full training set using its optimal k and evaluated once on the held-out test set. As this study used a static, pre-existing dataset rather than prospectively collected data, no data-collection time frame is applicable; all computational procedures were carried out within a single analysis session.

Standard classification performance metrics accuracy, precision, recall, and F1 score were used to compare the two models on the test set, alongside confusion matrices for descriptive comparison. Model selection (optimal k) was based on mean cross-validation accuracy across five folds, a standard and widely used statistical procedure that requires no additional citation (Widyadhana et al., 2021). The Fuzzy KNN algorithm itself, being a less commonly applied variant outside specialized literature, follows the formulation of (Uddin et al., 2022), with the fuzzifier parameter fixed at m = 2, consistent with common practice in the reviewed literature (Bian et al., 2022; Mailagaha Kumbure & Luukka, 2022).

This study is limited to a single, publicly available dataset representing a specific population (female patients of Pima Indian heritage), which constrains the generalizability of the findings to other populations or clinical settings. As a cross-sectional, pre-existing dataset, no temporal or longitudinal patterns could be assessed. The comparison was restricted to the original, untuned Fuzzy KNN formulation with a fixed fuzzifier and Euclidean distance metric; alternative distance metrics, fuzzifier values, or more advanced fuzzy KNN variants were outside the scope of this study. No formal statistical significance testing (e.g., paired t-tests) was conducted on the performance differences between models.

As this study is neither a clinical trial, qualitative study, systematic review, nor case report, none of the CONSORT, SRQR, PRISMA, or CARE guidelines directly apply. Given its nature as an analysis of an existing dataset, reporting in this manuscript was informed by the general principles of the STROBE guideline for observational studies (e.g., transparent reporting of data source, inclusion criteria, variables, and statistical methods), adapted appropriately for a computational/algorithmic comparison study rather than a traditional epidemiological observational design.

RESULTS AND DISCUSSION

Results:

Five-fold cross-validation on the training set was used to identify the optimal number of neighbors (k) for each model across a range of k = 3 to 21. The optimal k differed between the two methods: k = 15 for standard KNN and k = 3 for Fuzzy KNN. Despite this difference, cross-validation accuracy remained relatively stable across the tested k range for both methods, fluctuating narrowly between approximately 0.80 and 0.83, indicating that overall performance was not highly sensitive to the specific choice of k.

Table 2. Cross-validation accuracy range and optimal k for each model (k = 3, 4, ...,21)

Model	Optimal	CV Accuracy Range	CV Accuracy Range
KNN	15	0.8045 – 0.8258	0.8258
Fuzzy KNN (m = 2)	3	0.8045 – 0.8224	0.8224

When evaluated on the held-out test set (20% of the data, stratified split) using the respective optimal k for each model, Fuzzy KNN outperformed standard KNN across all four evaluations metrics: accuracy (0.779 vs. 0.773), precision (0.679 vs. 0.673), recall (0.704 vs. 0.685), and F1-score (0.691 vs. 0.679), as shown in Figure 2. The largest gap between the two models was observed in recall (+1.85 percentage points), suggesting that Fuzzy KNN was comparatively more effective at correctly identifying true diabetic cases.

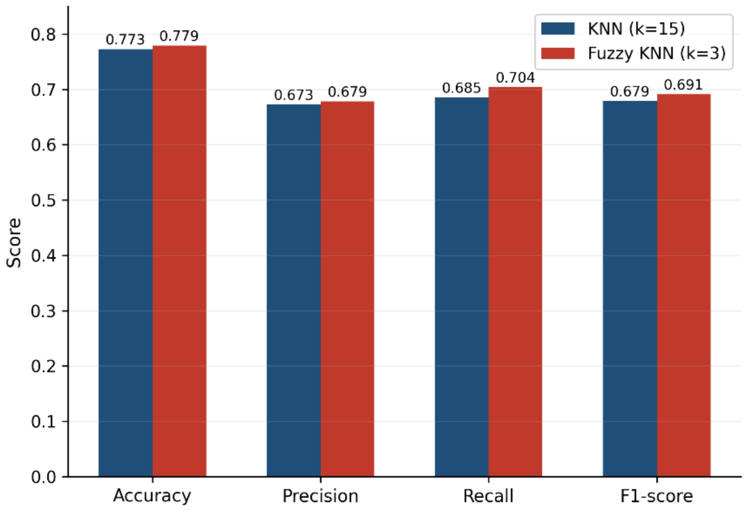

Figure 1. Comparison of test set performance (accuracy, precision, recall, F1-score) between KNN (k = 15) and Fuzzy KNN (k = 3)

Discussion:

These findings are consistent with the theoretical basis of Fuzzy KNN proposed by (Ren et al., 2026), in which each neighbor's contribution is weighted by its distance-based membership degree rather than counted equally in a majority vote. his weighting mechanism may contribute to the slightly higher recall observed for Fuzzy KNN, indicating that it identified a somewhat larger proportion of positive cases in the test set (X. Wu et al., 2026).

The difference in optimal k between the two models (15 for KNN versus 3 for Fuzzy KNN) is consistent with the different ways in which the two methods use neighborhood information. Standard KNN assigns equal voting influence to the selected neighbors, whereas Fuzzy KNN incorporates distance-based membership weighting. Thus, the distance weighting in Fuzzy KNN may help explain why a smaller neighborhood was sufficient to achieve its highest cross-validation accuracy. This finding is consistent with H2, although the observed difference in optimal k should be interpreted within the scope of the present dataset and experimental setting.

The improvement observed in the present study was modest, with differences of less than two percentage points across the four evaluations metrics. Reported higher performance for an enhanced Fuzzy KNN approach on the same dataset; however, their methodology included uncertainty-based membership, feature selection, and grid-search optimization. Therefore, the reported performance values should not be interpreted as a direct performance comparison between the two studies. Instead, the difference highlights the distinction in methodological scope: the present study focuses on the baseline effect of distance-based membership, whereas previous work has explored additional mechanisms to further optimize Fuzzy KNN performance.

Implications:

In the present experiment, the untuned Fuzzy KNN produced slightly higher values than standard KNN across the four evaluation metrics. This suggests that distance-based membership weighting can provide a modest performance advantage over crisp majority voting under the experimental conditions used in this study. The result may be relevant for researchers seeking a simple and interpretable baseline comparison before considering more complex Fuzzy KNN variants.

Research contribution:

This study contributes a transparent and reproducible baseline comparison between standard KNN and the original Fuzzy KNN formulation under a common experimental framework. Unlike studies that incorporate additional feature selection, hybridization, uncertainty-based membership, or optimization, the present analysis focuses on the basic difference between crisp majority voting and distance-based fuzzy membership. The systematic evaluation of k from 3 to 21 also provides an empirical comparison of the neighborhood sizes associated with the highest cross-validation accuracy for the two methods.

Limitations:

This study has several limitations. First, it relies on a single, relatively small dataset (768 instances) with a moderate class imbalance (65:35), which may limit the generalizability of the findings to other diabetes populations or clinical settings. Second, only the Euclidean distance metric and a fixed fuzzifier value (m = 2) were used for Fuzzy KNN; alternative distance metrics or fuzzifier values were not explored. Third, the performance differences observed across the four metrics were modest in magnitude and were not evaluated for statistical significance. Finally, the study did not incorporate class-imbalance handling techniques (e.g., oversampling) or feature selection, which may have influenced the absolute performance levels of both models.

Suggestions:

Future studies could extend this comparison by testing statistical significance of the observed differences (e.g., paired t-tests or McNemar's test across cross-validation folds), applying the comparison to larger and more diverse diabetes datasets, and exploring the effect of different fuzzifier values and distance metrics on Fuzzy KNN performance. Combining Fuzzy KNN with class-imbalance handling techniques or feature selection methods, and benchmarking it against more advanced fuzzy KNN variants (e.g., adaptive-k or uncertainty-membership tuning), would also help clarify how much of the performance gain achievable in more complex studies can be attributed to the core fuzzy-weighting mechanism versus additional optimization.

CONCLUSION

This study compared the performance of standard KNN and Fuzzy KNN in classifying diabetes status using the Pima Indians Diabetes Dataset and examined how the optimal number of neighbors (k) differed between the two methods. The results support H1 and H2 within the present experimental setting: Fuzzy KNN achieved slightly higher values than standard KNN across all four evaluations metrics on the held-out test set, and its optimal k (3) was smaller than that of standard KNN (15). The observed difference in optimal k is consistent with the different ways in which the two methods incorporate neighborhood information, with Fuzzy KNN using distance-based membership weighting rather than crisp majority voting. Overall, the findings suggest that Fuzzy KNN may provide a modest performance advantage over standard KNN for diabetes classification in the Pima Indians Diabetes Dataset, while maintaining a relatively simple and reproducible computational framework.

These findings provide a baseline for further investigation rather than definitive evidence of the general superiority of Fuzzy KNN. Future studies should evaluate the comparison using larger and more diverse diabetes datasets and assess the statistical significance of performance differences. Further investigation of alternative fuzzifier values, distance metrics, adaptive neighborhood selection, feature selection, and class-imbalance handling may also help determine whether the observed performance difference is maintained or improved under other experimental conditions.

ACKNOWLEDGMENT

The authors would like to thank the Faculty of Mathematics and Natural Sciences, Universitas Bengkulu, for the institutional support provided during this research. This study received no external funding and was conducted independently by the authors.

AUTHOR CONTRIBUTION STATEMENT

UR conceived and designed the study, conducted the data analysis and modeling, and wrote the original draft of the manuscript. SDK contributed to the literature review and data preprocessing. RW assisted in the experimental design and result interpretation. AN contributed to the visualization and validation of the results. OS reviewed and edited the manuscript. All authors read and approved the final manuscript.

REFERENCES

Addou, M., Mermri, E. B., & Gabli, M. (2026). Feature-Importance-Weighted Distance Metrics for Enhanced k-Nearest Neighbors Classification in Medical Diagnosis. *Human-Centric Intelligent Systems*, *6*(2), 283–300. https://doi.org/10.1007/s44230-026-00147-4

Akhtar, M., Quadir, A., Tanveer, M., & Arshad, M. (2024). Flexi-Fuzz least squares SVM for Alzheimer's diagnosis: Tackling noise, outliers, and class imbalance. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2410.14207

Amer, A. A., Ravana, S. D., & Habeeb, R. A. A. (2025). Effective k-nearest neighbor models for data classification enhancement. *Journal of Big Data*, *12*(1), 86. https://doi.org/10.1186/s40537-025-01137-2

Aria, A. O. H., Mulyana, D., Rifa'i, A. R., & Ikhsan, M. (2025). Performance Comparison of K-Nearest Neighbor, Decision Tree, and Support Vector Machine Algorithms for Diabetes Classification. *CoreID Journal*, *3*(3), 93–98. https://doi.org/10.60005/coreid.v3i3.139

Arowolo, M. O., Adebiyi, M. O., Adebiyi, A. A., & Olugbara, O. O. (2021). Optimized hybrid investigative based dimensionality reduction methods for malaria vector using KNN classifier. *Journal Of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00415-z

Azevedo, B. F., Rocha, A. M. A. C., & Pereira, A. I. (2024). Hybrid approaches to optimization and machine learning methods: a systematic literature review. *Machine Learning*, *113*(7), 4055–4097. https://doi.org/10.1007/s10994-023-06467-x

Baghdadi, N. A., Abdelaliem, S. M. F., Malki, A., Gad, I., Ewis, A. A., & Atlam, E.-S. (2023). Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. *Journal Of Big Data*, *10*(1). https://doi.org/10.1186/s40537-023-00817-1

Bian, Z., Vong, C. M., Wong, P. K., & Wang, S. (2022). Fuzzy KNN Method With Adaptive Nearest Neighbors. *IEEE Transactions on Cybernetics*, *52*(6), 5380–5393. https://doi.org/10.1109/TCYB.2020.3031610

Bonmati, E., Hu, Y., Grimwood, A., Johnson, G., Goodchild, G., Keane, M. G., Gurusamy, K. S., Davidson, B. R., Clarkson, M. J., Pereira, S. P., & Barratt, D. C. (2021). Voice-Assisted Image Labeling for Endoscopic Ultrasound Classification Using Neural Networks. *ArXiv (Cornell University)*, *41*(6), 1311–1319. https://doi.org/10.1109/tmi.2021.3139023

Calster, B. Van, Steyerberg, E. W., Wynants, L., & Smeden, M. van. (2023). There is no such thing as a validated prediction model. *BMC Medicine*, *21*(1), 70. https://doi.org/10.1186/s12916-023-02779-w

Chandrasekaran, S., Dutt, V., Vyas, N., & Anand, A. (2023). Fuzzy KNN Implementation for Early Parkinson's Disease Prediction. *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, 896–901.https://doi.org/10.1109/ICCMC56507.2023.10083522

Choubey, D. K., Kumar, M., Shukla, V., Tripathi, S., & Dhandhania, V. K. (2020). Comparative Analysis of Classification Methods with PCA and LDA for Diabetes. *Current Diabetes Reviews*, *16*(8), 833–850. https://doi.org/10.2174/1573399816666200123124008

Dinesh, P., Vickram, A. S., & Kalyanasundaram, P. (2024). *Medical image prediction for diagnosis of breast cancer disease comparing the machine learning algorithms: SVM, KNN, logistic regression, random forest and decision tree to measure accuracy*. 020140. https://doi.org/10.1063/5.0203746

Fadaei-Kermani, E., & Ghaeini-Hessaroeyeh, M. (2020). Fuzzy nearest neighbor approach for drought monitoring and assessment. *Applied Water Science*, *10*(6). https://doi.org/10.1007/s13201-020-01212-4

Fan, C. (2025). Optimization and performance evaluation of machine learning classifiers for predicting construction quality and schedule. *Automation in Construction*, *179*, 106470. https://doi.org/10.1016/j.autcon.2025.106470

González-Martínez, S., García, S., Li, S.-T., John, R., & Herrera, F. (2020). Fuzzy k-Nearest Neighbors with monotonicity constraints: Moving towards\n the robustness of monotonic noise. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2003.02601

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, *11*(1), 113. https://doi.org/10.1186/s40537-024-00973-y

Hameed, E. M., & Joshi, H. (2024). Improving Diabetes Prediction by Selecting Optimal K and Distance Measures in KNN Classifier. *Journal of Techniques*, *6*(3), 19–25. https://doi.org/10.51173/jt.v6i3.2587

Kaur, J., & Khehra, B. S. (2022). Fuzzy Logic and Hybrid based Approaches for the Risk of Heart Disease Detection: State-of-the-Art Review. *Journal of The Institution of Engineers (India): Series B*, *103*(2), 681–697. https://doi.org/10.1007/s40031-021-00644-z

Lahmiri, S. (2023). Integrating convolutional neural networks, kNN, and Bayesian optimization for efficient diagnosis of Alzheimer's disease in magnetic resonance images. *Biomedical Signal Processing and Control*, *80*, 104375. https://doi.org/10.1016/j.bspc.2022.104375

Li, Y., Zhao, D., Xu, Z., Heidari, A. A., Chen, H., Jiang, X., Liu, Z., Wang, M., Zhou, Q., & Xu, S. (2023). bSRWPSO-FKNN: A boosted PSO with fuzzy K-nearest neighbor classifier for predicting atopic dermatitis disease. *Frontiers in Neuroinformatics*, *16*, 1063048. https://doi.org/10.3389/fninf.2022.1063048

Lohani, B. P., Dagur, A., & Shukla, D. K. (2025). Synergistic ensemble classification framework: utilizing a soft voting algorithm for enhanced prediction and diagnosis of diabetes mellitus. *Indonesian Journal of Electrical Engineering and Computer Science*, *37*(3), 1945. https://doi.org/10.11591/ijeecs.v37.i3.pp1945-1953

Lohrmann, C., Lohrmann, A., & Kumbure, M. M. (2025). On the benefit of feature selection and ensemble feature selection for fuzzy k-nearest neighbor classification. *Applied Soft Computing*, *171*, 112784. https://doi.org/10.1016/j.asoc.2025.112784

Mailagaha Kumbure, M., & Luukka, P. (2022). A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance. *Granular Computing*, *7*(3), 657–671. https://doi.org/10.1007/s41066-021-00288-w

Mailagaha Kumbure, M., Luukka, P., & Collan, M. (2020). A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean. *Pattern Recognition Letters*, *140*, 172–178. https://doi.org/10.1016/j.patrec.2020.10.005

Mayawi, M., & Subanar, S. (2025). Parameter Independent Fuzzy Weighted K-Nearest Neighbor. *Media Statistika*, *17*(2), 162–172. https://doi.org/10.14710/medstat.17.2.162-172

Pikula, A., Gulati, M., Bonnet, J. P., Ibrahim, S., Chamoun, S., Freeman, A. M., & Reddy, K. (2024). Promise of Lifestyle Medicine for Heart Disease, Diabetes Mellitus, and Cerebrovascular Diseases. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, *8*(2), 151–165. https://doi.org/10.1016/j.mayocpiqo.2023.11.005

Ren, J., Carroll, H., McCarthy, K., Allen, J., Tam, J., Philip, J., Mehta, M., & Clark, D. W. (2026). A kNN based machine learning approach to automating causality assessment of adverse events. *Scientific Reports*, *16*(1). https://doi.org/10.1038/s41598-026-40267-2

Sachdeva, R. K., Bathla, P., Rani, P., Lamba, R., Ghantasala, P., & Nassar, I. F. (2025). No TitleA Novel K Nearest Neighbor Classifier for Lung Cancer Disease Diagnosis. *Chemotherapy: Open Access*, *13*(1), 1–9. https://doi.org/10.35248/2167-7700.25.13.231

Salem, H., Shams, M. Y., Elzeki, O. M., Abd Elfattah, M., F. Al-Amri, J., & Elnazer, S. (2022). Fine-Tuning Fuzzy KNN Classifier Based on Uncertainty Membership for the Medical Diagnosis of Diabetes. *Applied Sciences*, *12*(3), 950. https://doi.org/10.3390/app12030950

Siyar, S., Azarnoush, H., Rashidi, S., Winkler-Schwartz, A., Bissonnette, V., Ponnudurai, N., & Maestro, R. F. Del. (2020). Machine learning distinguishes neurosurgical skill levels in a virtual reality tumor resection task. *ArXiv (Cornell University)*, *58*(6), 1357–1367. https://doi.org/10.1007/s11517-020-02155-3

Stefane Souza, V., & Araújo Lima, D. (2024). Cardiac Disease Diagnosis Using K-Nearest Neighbor Algorithm: A Study on Heart Failure Clinical Records Dataset. *Artificial Intelligence and Applications*, *3*(1), 56–71. https://doi.org/10.47852/bonviewAIA42022045

Suárez, J. L., González-Almagro, G., García, S., & Herrera, F. (2024). Metric learning for monotonic classification: turning the space up to the limits of monotonicity. *Applied Intelligence*, *54*(5), 4443–4466. https://doi.org/10.1007/s10489-024-05371-8

Suyal, M., & Goyal, P. (2022). A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning. *International Journal of Engineering Trends and Technology*, *70*(7), 43–48. https://doi.org/10.14445/22315381/IJETT-V70I7P205

Syriopoulos, P. K., Kalampalikis, N. G., Kotsiantis, S. B., & Vrahatis, M. N. (2025). kNN Classification: a review. *Annals of Mathematics and Artificial Intelligence*, *93*(1), 43–75. https://doi.org/10.1007/s10472-023-09882-x

Szczuko, P., Kurowski, A., Odya, P., Czyżewski, A., Kostek, B., Graff, B., & Narkiewicz, K. (2021). Mining Knowledge of Respiratory Rate Quantification and Abnormal Pattern Prediction. *Cognitive Computation*, *14*(6), 2120–2140. https://doi.org/10.1007/s12559-021-09908-8

Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, *10*(1–2), 1–10. https://doi.org/10.1049/htl2.12039

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, *12*(1), 6256. https://doi.org/10.1038/s41598-022-10358-x

Vommi, A. M., & Krishna, B. T. (2024). An equilibrium optimizer-based parameter independent fuzzy kNN classifier for classification of medical datasets. *Soft Computing*, *28*(20), 11757–11765. https://doi.org/10.1007/s00500-024-09941-3

Widyadhana, A., Putra, C. B. P., Indraswari, R., & Arifin, A. Z. (2021). A Bonferroni Mean Based Fuzzy K Nearest Centroid Neighbor Classifier. *Jurnal Ilmu Komputer Dan Informasi*, *14*(1), 65–71. https://doi.org/10.21609/jiki.v14i1.959

Wu, W., Keller, J. M., Skubic, M., Popescu, M., & Lane, K. R. (2021). Early Detection of Health Changes in the Elderly Using In-Home Multi-Sensor Data Streams. *ACM Transactions on Computing for Healthcare*, *2*(3), 1–23. https://doi.org/10.1145/3448671

Wu, X., Ding, Y., Wang, L., Ding, D., & Zhang, H. (2026). A novel Neutrosophic Transformer for handling imbalanced data. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1007/s44443-026-00664-z

Zhang, S. (2022). Challenges in KNN Classification. *IEEE Transactions on Knowledge and Data Engineering*, *34*(10), 4663–4675. https://doi.org/10.1109/TKDE.2021.3049250