CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (2) (2026), Pages 1544–1561
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Diabetes Classification Using SMOTE-Based Intuitionistic Fuzzy K-NN with Weighted HIAD

Tio Valent Novi Yanti Sianturi and Raden Sulaiman*

Department of Mathematics, Faculty of Mathematics and Natural Sciences, State University of Surabaya, Surabaya, Indonesia.

Abstract

Diabetes risk classification is challenged by class imbalance and diagnostic uncertainty. This study investigates an Intuitionistic Fuzzy Set framework integrated with a proposed Hausdorff-Inspired Attribute Distance within the K-Nearest Neighbor algorithm. The framework models patient profiles using membership, non-membership, and hesitation components to capture clinical ambiguity. We evaluated this approach on the Pima Indians Diabetes dataset using Stratified 5-Fold Cross-Validation, incorporating class balancing and clinically informed feature weighting. Results showed that while a conventional Euclidean-based model achieved the highest accuracy (73.96%) and data balancing maximized sensitivity (77.61%), the intuitionistic fuzzy configurations achieved the highest specificity (89.80%), indicating a conservative uncertainty-aware classification behavior. The proposed distance measure yielded classification performance highly comparable to that of the conventional Hamming distance, providing an alternative maximum-deviation-based similarity formulation within the intuitionistic fuzzy framework. Furthermore, combining class balancing with feature weighting contributed to a more balanced sensitivity–specificity profile. Overall, the findings demonstrate how uncertainty representation, feature weighting, class balancing, and neighborhood configuration influence classification outcomes, highlighting the potential of intuitionistic fuzzy approaches for uncertainty-aware medical decision-support systems.

Keywords: Class Imbalance; Diabetes Risk Classification; Hausdorff-Inspired Attribute Distance; Intuitionistic Fuzzy Set; K-Nearest Neighbor.

1. Introduction

Diabetes Mellitus (DM) is a chronic metabolic disorder characterized by persistent hyperglycemia resulting from impaired insulin secretion, insulin resistance, or both [1]. The prevalence of diabetes continues to increase globally and has become one of the major challenges in modern healthcare systems. According to the International Diabetes Federation (IDF), approximately 537 million adults worldwide are currently living with diabetes, and this number is projected to increase to 783 million by 2045 [2]. This rapid growth highlights the importance of developing reliable approaches for early detection of diabetes risk and clinical decision support systems[3].

Early diabetes identification is essential because delayed diagnosis may lead to severe complications, including cardiovascular disease, kidney failure, neuropathy, and visual impairment.

*Corresponding author. E-mail: radensulaiman@unesa.ac.id

Received: April 15, 2026 Revised: May 22, 2026 Accepted: June 25, 2026
DOI: https://doi.org/10.18860/cauchy.v11i2.41935

In clinical practice, diabetes diagnosis involves multiple physiological indicators such as blood glucose level, Body Mass Index (BMI), blood pressure, insulin concentration, age, and heredity. However, these clinical attributes frequently exhibit overlapping characteristics between diabetic and non-diabetic individuals, particularly near diagnostic threshold regions defined by organizations such as the World Health Organization (WHO) and the American Diabetes Association (ADA) [4]. Consequently, diabetes risk assessment inherently involves uncertainty and transitional conditions that may be difficult to model with strict binary decision boundaries [5–7].

Various machine learning methods have been widely investigated for diabetes classification and prediction. Several studies reported that ensemble learning and deep learning approaches are capable of achieving competitive classification performance on medical datasets [8, 9]. Nevertheless, many conventional classification models still rely on deterministic numerical representations and standard geometric distance measures such as Euclidean distance. In distance-based classifiers such as K-Nearest Neighbor (K-NN), the Euclidean metric may smooth localized attribute variations during similarity computation, which could reduce sensitivity toward subtle but clinically important differences among patient profiles near diagnostic boundaries [10, 11].

To address uncertainty in medical data, Fuzzy Logic-based approaches have also been explored in diabetes classification studies. Classical fuzzy sets enable gradual membership representation rather than rigid binary categorization, making them more suitable for modeling ambiguous medical conditions [5, 6, 12]. However, most existing fuzzy-based frameworks only consider membership information and do not explicitly represent hesitation or uncertainty margins during the similarity evaluation process [13]. As a result, uncertainty near overlapping diagnostic boundaries may not be adequately preserved, potentially affecting classification behavior when handling imbalanced clinical distributions.

These limitations indicate the need for a classification framework capable of explicitly representing uncertainty while preserving localized attribute deviations in clinical patient data. One promising approach is the Intuitionistic Fuzzy Set (IFS), introduced by Atanassov [14]. Unlike classical fuzzy sets, IFS simultaneously represents membership, non-membership, and hesitation degrees, thereby providing a more flexible mathematical representation for handling uncertain and overlapping clinical information [15, 16].

Motivated by this research gap, this study proposes an Intuitionistic Fuzzy Set (IFS)-based diabetes classification framework integrated with a Hausdorff-Inspired Attribute Distance (HIAD) metric within the K-Nearest Neighbor (K-NN) algorithm. The proposed HIAD formulation adopts the maximum-deviation principle inspired by the Hausdorff distance and adapts it into an attribute-wise similarity computation mechanism for structured tabular clinical data [17, 18]. The proposed formulation emphasizes the largest deviation among the membership, non-membership, and hesitation components during neighborhood similarity evaluation, providing an alternative similarity-computation strategy within the intuitionistic fuzzy space for comparing patient profiles under uncertainty. Rather than assuming an inherent performance advantage, the objective is to investigate whether the maximum-deviation principle offers a different perspective on modeling similarity relationships in uncertainty-aware medical classification tasks. Furthermore, the proposed HIAD formulation is evaluated alongside the conventional Hamming distance to investigate whether alternative similarity-aggregation mechanisms lead to different classification behaviors within the same IFS representation.

To investigate the influence of class balancing and feature importance within the proposed framework, this study incorporates the Synthetic Minority Over-sampling Technique (SMOTE) [19] alongside a clinically informed feature-weighting strategy. Rather than aiming for single-metric maximization, the framework focuses on examining the trade-offs between diabetic case detection and healthy-patient identification under different classification settings. A comprehensive ablation analysis is conducted across multiple classification scenarios to systematically evaluate the individual and combined effects of uncertainty representation, adapted similarity measurement, class balancing, and feature weighting within the proposed framework.

The main contributions of this study can be summarized as follows:

1. Development of an Intuitionistic Fuzzy Set (IFS)-based diabetes classification framework capable of explicitly representing membership, non-membership, and hesitation information within clinical patient profiles to represent uncertainty in diabetes-risk classification.
2. Proposal of a Hausdorff-Inspired Attribute Distance (HIAD) formulation that adapts the maximum-deviation principle into an attribute-wise similarity computation mechanism within the Intuitionistic Fuzzy Set framework, providing an alternative approach for evaluating neighborhood relationships under uncertainty.
3. Implementation of a hybrid framework combining IFS-HIAD, SMOTE, and clinically informed feature weighting, accompanied by a comprehensive ablation analysis to examine how uncertainty representation, class balancing, and feature weighting influence sensitivity–specificity trade-offs across varying neighborhood configurations.

The remainder of this paper is organized as follows. Section 2 presents the research methodology, including the theoretical framework of Intuitionistic Fuzzy Sets (IFS), the proposed HIAD formulation, and the classification procedure. Section 3 discusses the experimental results and performance analysis across multiple classification scenarios. Finally, Section 4 concludes the study and outlines potential directions for future research.

2. Methods

This study employs a computational approach for Type 2 *Diabetes Mellitus* (DM) classification. The proposed framework integrates *Intuitionistic Fuzzy Sets* (IFS), a *Hausdorff-Inspired Attribute Distance* (HIAD) formulation, and the *K-Nearest Neighbor* (K-NN) algorithm to represent uncertainty and ambiguity commonly encountered in clinical data.

2.1. Dataset and Data Preprocessing

The experiments were conducted using the *Pima Indians Diabetes Dataset*, publicly available through the *UCI Machine Learning Repository* and *Kaggle* [20, 21]. The dataset contains 768 observations with eight numerical predictor attributes, namely *Pregnancies*, *Glucose*, *Blood Pressure*, *Skin Thickness*, *Insulin*, *Body Mass Index* (BMI), *Diabetes Pedigree Function*, and *Age*. The target variable is binary, where class 1 represents diabetic patients and class 0 represents non-diabetic patients.

To ensure a leakage-free evaluation, all preprocessing procedures were performed within each training fold of a *Stratified 5-Fold Cross-Validation* framework and subsequently applied to the corresponding validation fold. The same fold partitions were used consistently across all experimental scenarios to ensure a fair comparison.

1. **Missing Value Imputation**
   Zero values in *Glucose*, *Blood Pressure*, *Skin Thickness*, *Insulin*, and *BMI* were treated as biologically implausible and imputed using the median estimated exclusively from the training fold.
2. **Min-Max Normalization**
   All features were normalized into the interval $[0, 1]$ using:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

   The normalization parameters were fitted on the training fold and subsequently applied to the validation fold. Since fuzzification was performed in the normalized feature space, all fuzzy threshold parameters (a, b, c) were transformed using the same normalization parameters to maintain consistency between the normalized data and the fuzzy membership functions [22].

3. **Data Balancing Using SMOTE**
   To address class imbalance, the *Synthetic Minority Over-sampling Technique* (SMOTE) [19] was applied exclusively to the normalized training data within each fold. No synthetic samples were generated from or applied to the validation data.

Fig. 1 presents the overall workflow of the proposed framework, including preprocessing, IFS transformation, distance computation, and K-NN classification.

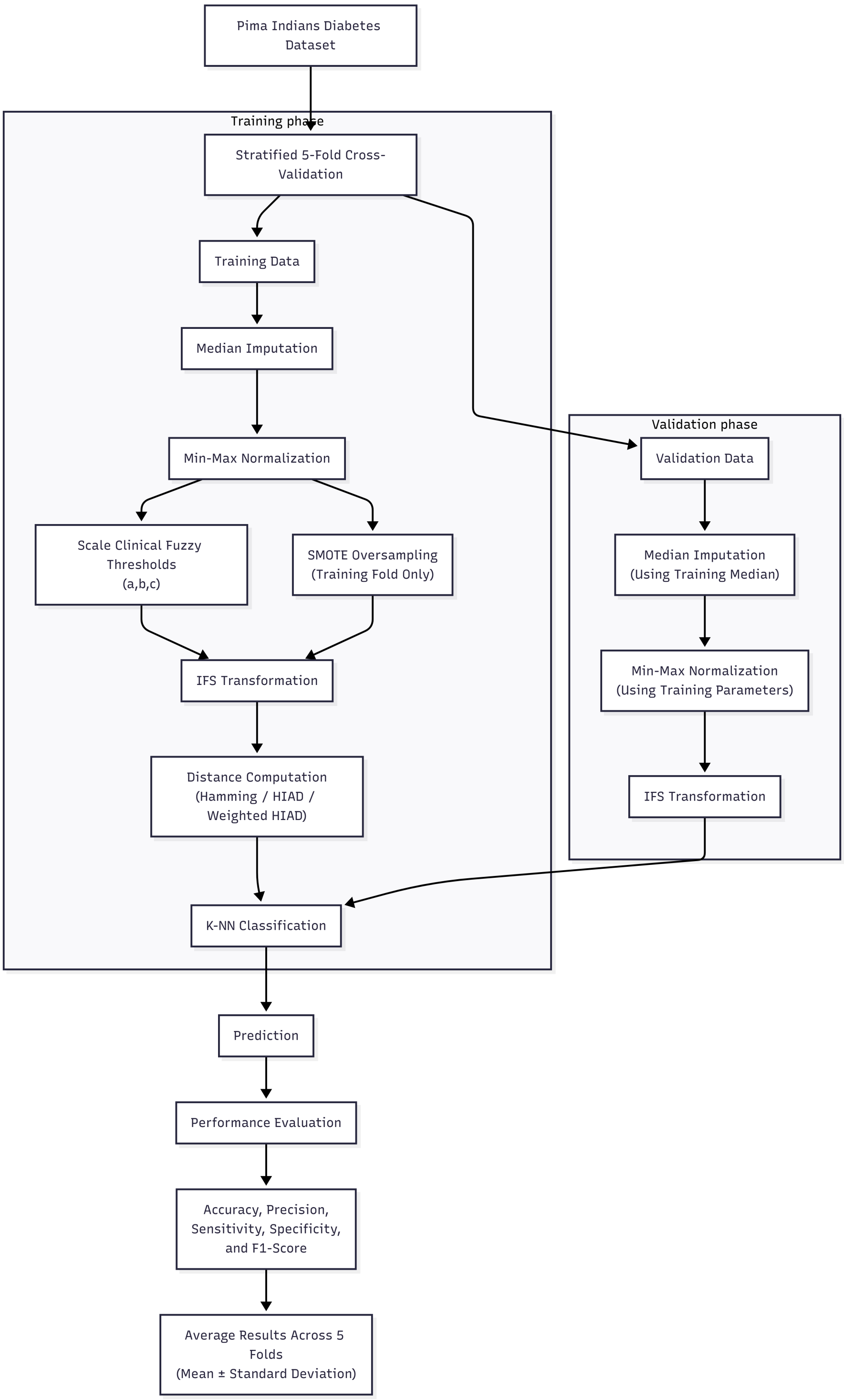

Fig. 1: Overall workflow of the proposed diabetes classification framework.

2.2. Determination of Clinical Fuzzy Parameters

The fuzzy membership parameters were determined using a hybrid clinically informed and data-driven approach. Each attribute was represented by a triangular membership function (a, b, c) corresponding to *low-risk*, *moderate-risk*, and *high-risk* regions.

For clinically relevant attributes such as *Glucose*, *Blood Pressure*, and *BMI*, the parameters were constructed using reference ranges reported by the American Diabetes Association (ADA) [4] and the World Health Organization (WHO) [1]. For attributes without universally accepted clinical thresholds, including *Pregnancies*, *Skin Thickness*, *Insulin*, *Diabetes Pedigree Function* (DPF), and *Age*, dataset-informed modeling assumptions were adopted to preserve the distributional characteristics of the clinical data within the experimental setting. The resulting fuzzy parameters are summarized in Table 1.

Table 1: Hybrid Clinical and Data-Driven Fuzzy Parameters for IFS Transformation

No.	Patient Attribute	Lower (a) *(Low-Risk)*	Core (b) *(Mod-Risk)*	Upper (c) *(High-Risk)*	Basis / Reference
1	Pregnancies (Times)	1	3	5	Dataset-driven assumption
2	Glucose (mg/dL)	140	170	200	ADA-informed threshold
3	Blood Pressure (mmHg)	80	85	90	Clinically informed modeling
4	Skin Thickness (mm)	20	25	30	Dataset-driven assumption
5	Insulin (μU/mL)	140	170	200	Dataset-driven assumption
6	BMI (kg/m^2)	24.9	27.5	30	WHO obesity classification
7	DPF (Score)	0.30	0.55	0.80	Dataset-driven assumption
8	Age (Years)	34	40	45	Dataset-driven assumption

The parameters in Table 1 should be interpreted as gradual transition regions rather than strict diagnostic cut-off values. For clinically informed attributes, the triangular membership functions were constructed around risk-related reference ranges, whereas the remaining attributes were modeled using dataset-informed assumptions. This distinction is maintained consistently throughout the manuscript to ensure methodological transparency regarding the basis of each threshold.

2.3. Intuitionistic Fuzzy Set (IFS) Representation

Intuitionistic Fuzzy Sets (IFS), introduced by Atanassov [14], extend classical fuzzy sets by incorporating three complementary components: membership (μ), non-membership (ν), and hesitation (π). This representation provides an additional mechanism for modeling uncertainty and ambiguity in clinical data, where patient characteristics may overlap across diagnostic categories [16, 23].

Definition 1. Let X denote the universe of discourse. An Intuitionistic Fuzzy Set (IFS) A in X is defined as:

$$A = \{\langle x, \mu_A(x), \nu_A(x)\rangle \mid x \in X\}$$

where $\mu_A(x) : X \to [0, 1]$ denotes the membership degree and $\nu_A(x) : X \to [0, 1]$ denotes the non-membership degree, subject to:

$$0 \le \mu_A(x) + \nu_A(x) \le 1$$

The hesitation degree is defined as:

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x)$$

where $\pi_A(x)$ represents the hesitation degree associated with the observation $x \in X$.

2.4. Membership Function and Parameter λ

To transform normalized continuous clinical attributes into the Intuitionistic Fuzzy Set (IFS) space, this study employs a triangular membership function parameterized by the fuzzy thresholds (a, b, c) established in Table 1. The triangular membership function was selected due to its computational simplicity, interpretability, and effectiveness in modeling gradual transitions between clinical risk categories [24].

Fuzzification was performed after Min-Max normalization to ensure consistency between the normalized feature space and the fuzzy threshold parameters. For a patient sample i with respect to attribute j, the membership degree $\mu(x_{ij})$ is defined as:

$$\mu(x_{ij}; a_j, b_j, c_j) = \begin{cases} 0, & \text{if } x_{ij} \le a_j \text{ or } x_{ij} \ge c_j \\ \dfrac{x_{ij} - a_j}{b_j - a_j}, & \text{if } a_j < x_{ij} \le b_j \\ \dfrac{c_j - x_{ij}}{c_j - b_j}, & \text{if } b_j < x_{ij} < c_j \end{cases}$$

The degree of non-membership is computed using a controlled complementary formulation:

$$\nu(x_{ij}) = (1 - \mu(x_{ij}))\lambda \tag{1}$$

where $\lambda \in [0, 1]$ denotes the non-membership control parameter. The hesitation degree is subsequently calculated as:

$$\pi(x_{ij}) = 1 - \mu(x_{ij}) - \nu(x_{ij}) \tag{2}$$

Substituting Eq. (1) into Eq. (2) yields:

$$\pi(x_{ij}) = (1 - \mu(x_{ij}))(1 - \lambda) \tag{3}$$

Eq. (3) shows that λ controls the distribution of uncertainty between the non-membership and hesitation components. Larger values increase the non-membership degree while reducing hesitation, whereas smaller values preserve a greater proportion of uncertainty within the IFS representation.

In this study, $\lambda = 0.8$ was adopted. A localized sensitivity analysis using $\lambda \in \{0.5, 0.6, 0.7, 0.8\}$ produced highly similar classification results, suggesting that moderate variations in λ did not substantially alter the induced neighborhood structure. This observation may further indicate that the chosen IFS construction does not strongly differentiate neighborhood relationships within the tested parameter range. Detailed results are presented in Section 3. Fig. 2 illustrates the resulting membership (μ), non-membership (ν), and hesitation (π) functions.

2.5. Hausdorff-Inspired Attribute Distance (HIAD)

To measure the proximity between patient profiles within the Intuitionistic Fuzzy Set (IFS) space, this study introduces the *Hausdorff-Inspired Attribute Distance* (HIAD). The formulation is inspired by the classical Hausdorff metric, which preserves the maximum deviation between two sets by identifying the largest local discrepancy between their corresponding elements [10, 17].

In its classical form, the Hausdorff distance between two non-empty sets X and Y is defined as:

$$d_H(X, Y) = \max\left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(y, x) \right\}$$

where $d(x, y)$ denotes the distance between elements x and y.

Although the classical Hausdorff distance is effective for geometric point-set comparisons, it is not directly applicable to structured tabular clinical data represented as fixed attribute vectors. Therefore, this study adapts the maximum-deviation principle into an attribute-wise

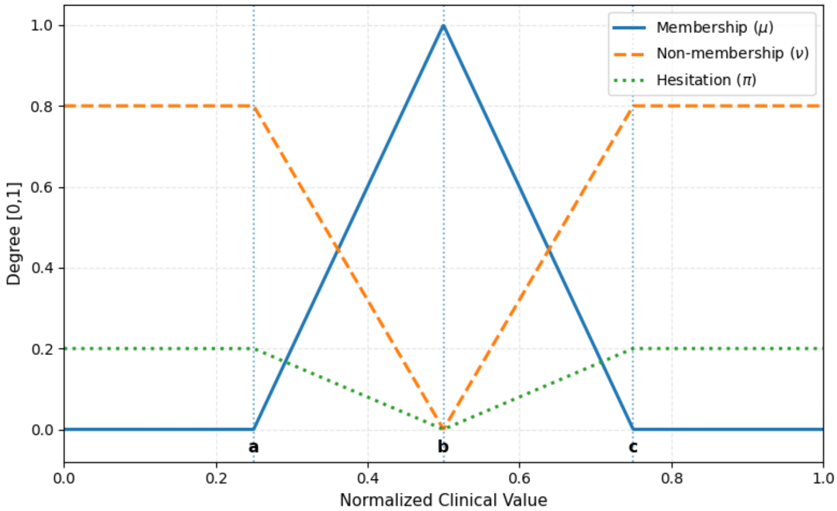

Fig. 2: Schematic visualization of the Intuitionistic Fuzzy Set (IFS) representation showing the membership (μ), non-membership (ν), and hesitation (π) functions generated from a triangular membership function with parameters a, b, and c using $\lambda = 0.8$.

similarity mechanism within the IFS framework. Consequently, HIAD should be interpreted as a Hausdorff-inspired distance measure rather than a direct implementation of the classical Hausdorff distance.

Formally, the unweighted HIAD between two patient profiles A and B is defined as:

$$D_{\text{HIAD}}(A, B) = \frac{1}{n}\sum_{i=1}^{n} \max\Big(|\mu_A(x_i) - \mu_B(x_i)|\,, |\nu_A(x_i) - \nu_B(x_i)|\,, |\pi_A(x_i) - \pi_B(x_i)| \Big)$$

where:

- n denotes the total number of clinical attributes;
- $\mu_A(x_i)$ and $\mu_B(x_i)$ represent the membership degrees of attribute i for patient profiles A and B;
- $\nu_A(x_i)$ and $\nu_B(x_i)$ represent the corresponding non-membership degrees; and
- $\pi_A(x_i)$ and $\pi_B(x_i)$ represent the corresponding hesitation degrees.

The maximum operator preserves the dominant discrepancy among the membership, non-membership, and hesitation components, thereby maintaining sensitivity to localized clinical abnormalities while reducing the smoothing effect commonly associated with global aggregation schemes such as Euclidean distance.

From a clinical perspective, a substantial deviation in a single clinically relevant attribute may indicate a different diabetes risk profile, even when the remaining attributes appear similar. By evaluating each attribute independently before aggregation, HIAD preserves such localized differences while still providing a global assessment of patient similarity.

While the unweighted HIAD treats all clinical attributes equally, different medical indicators do not contribute equally to diabetes-risk assessment. Therefore, a weighted extension of HIAD is introduced in the next subsection to incorporate the relative clinical importance of each attribute.

2.6. Weighted HIAD

While the unweighted HIAD formulation treats all clinical attributes equally, different medical indicators do not contribute equally to diabetes risk assessment. Therefore, this study extends the HIAD framework by incorporating attribute-specific weights to reflect the relative clinical importance of each feature.

The weighted HIAD between patient profiles A and B is defined as:

$$D_{\text{WHIAD}}(A, B) = \sum_{i=1}^{n} w_i \cdot \max\Big(|\mu_A(x_i) - \mu_B(x_i)|, |\nu_A(x_i) - \nu_B(x_i)|, |\pi_A(x_i) - \pi_B(x_i)| \Big)$$

subject to the normalization constraints:

$$\sum_{i=1}^{n} w_i = 1, \qquad w_i \geq 0 \tag{4}$$

where w_i denotes the normalized importance weight associated with the i-th attribute. To satisfy Eq. (4), each weight was computed as:

$$w_i = \frac{\omega_i}{\sum_{j=1}^{n} \omega_j}$$

where ω_i represents the raw clinically informed importance score assigned to attribute i.

The weighting configuration was determined using a clinically informed heuristic strategy rather than directly adopting numerical values from clinical guidelines. Attributes commonly recognized as stronger diabetes-related risk indicators were assigned larger importance scores, whereas supporting indicators received lower weights.

Table 2: Clinically Informed Feature Weights Used in the Proposed Weighted HIAD

Attribute	Raw Weight (ω_i)	Normalized Weight (w_i)
Pregnancies	1.0	0.0833
Glucose	3.5	0.2917
Blood Pressure	1.0	0.0833
Skin Thickness	0.5	0.0417
Insulin	1.0	0.0833
BMI	2.0	0.1667
DPF	1.5	0.1250
Age	1.5	0.1250
Total	$\sum \omega_i = 12.0$	$\sum w_i = 1.0000$

As shown in Table 2, Glucose received the largest normalized weight (0.2917), followed by BMI (0.1667). Age and Diabetes Pedigree Function (DPF) were assigned moderate weights, whereas the remaining attributes received lower weights as supporting indicators. These values represent modeling assumptions informed by clinical knowledge rather than quantitative weights derived directly from clinical guidelines.

The weighted HIAD formulation incorporates clinically informed feature importance into similarity computation, enabling the Intuitionistic Fuzzy K-NN classifier to place greater emphasis on medically relevant attributes while preserving the uncertainty representation provided by the IFS framework.

2.7. Classification via Intuitionistic Fuzzy K-Nearest Neighbor

The final classification stage employs a K-Nearest Neighbor (K-NN) framework operating in the Intuitionistic Fuzzy Set (IFS) space. Unlike conventional K-NN, which relies on standard geometric distances, the proposed framework evaluates neighborhood relationships using intuitionistic fuzzy representations consisting of membership (μ), non-membership (ν), and hesitation (π) components.

The classification procedure consists of four sequential steps:

1. **Intuitionistic Fuzzification:** Transforming normalized clinical attributes into IFS representations.

2. **Distance Computation:** Calculating pairwise distances using the selected metric (Euclidean, IFS-Hamming, IFS-HIAD, or Weighted HIAD).
3. **Neighborhood Selection:** Identifying the K nearest neighbors based on ascending distance values.
4. **Majority Voting:** Assigning the final diagnostic label according to the most frequent class among the selected neighbors.

To evaluate the influence of neighborhood size, the following odd-valued configurations were investigated:

$$K \in \{3, 5, 7, 9\} \tag{5}$$

Odd values were selected to reduce tie conditions during binary classification. All experiments were evaluated using a Stratified 5-Fold Cross-Validation protocol to preserve class distributions across folds and obtain robust performance estimates.

2.8. Experimental Ablation Design

To evaluate the contribution of each methodological component, an ablation-based experimental design consisting of six classification scenarios was employed:

- **Scenario 1:** K-NN (Euclidean);
- **Scenario 2:** K-NN + SMOTE;
- **Scenario 3:** IFS-Hamming;
- **Scenario 4:** IFS-HIAD;
- **Scenario 5:** IFS-HIAD + SMOTE;
- **Scenario 6:** Weighted HIAD + SMOTE.

Scenario 1 serves as the baseline model, whereas Scenarios 2–6 progressively introduce class balancing, intuitionistic fuzzy representation, the Hausdorff-Inspired Attribute Distance (HIAD), and clinically informed feature weighting. This design enables the systematic examination of the individual and combined effects of these components.

To ensure a fair comparison, all scenarios were evaluated using identical Stratified 5-Fold Cross-Validation partitions and the neighborhood configurations defined in Eq. (5). All preprocessing procedures, including median imputation, Min-Max normalization, fuzzy-threshold scaling, and SMOTE oversampling, were performed exclusively on the training folds and subsequently applied to the corresponding validation folds, thereby maintaining a leakage-free evaluation protocol.

2.9. Model Evaluation

Classification performance was evaluated using confusion matrix-based metrics, namely Accuracy, Precision, Sensitivity, Specificity, and F1-Score. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The evaluation metrics are defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

Accuracy measures overall classification performance; Precision evaluates the reliability of positive predictions; Sensitivity the detection of diabetic patients; and Specificity measures the correct identification of non-diabetic individuals. The F1-Score provides a harmonic balance between Precision and Sensitivity and is particularly useful for imbalanced datasets.

Given the clinical importance of diabetic case detection, particular attention was paid to Sensitivity alongside overall accuracy. All scenarios were evaluated using the Stratified 5-Fold Cross-Validation protocol described in Section 2.7. Reported results correspond to the mean and standard deviation across the five validation folds. All experiments were implemented in Python 3.10 using `NumPy`, `Pandas`, `Scikit-learn`, and `Imbalanced-learn`.

3. Results and Discussion

The experimental results were obtained using the leakage-free Stratified 5-Fold Cross-Validation protocol described in Section 2.7. All reported values correspond to the mean and standard deviation across the five validation folds. The analysis evaluates the effects of class balancing, intuitionistic fuzzy representation, distance metric selection, and clinically informed feature weighting on diabetes classification performance.

3.1. Performance Comparison Across Classification Scenarios

Table 3 summarizes the classification results obtained from the six experimental scenarios across four neighborhood configurations ($K \in \{3, 5, 7, 9\}$). The reported metrics include Accuracy, Precision, Sensitivity, Specificity, and F1-Score, providing a comprehensive basis for evaluating the effects of SMOTE, Intuitionistic Fuzzy Set (IFS) representation, Hausdorff-Inspired Attribute Distance (HIAD), and clinically informed feature weighting.

Fig. 3 presents the average classification performance across all evaluated neighborhood configurations.

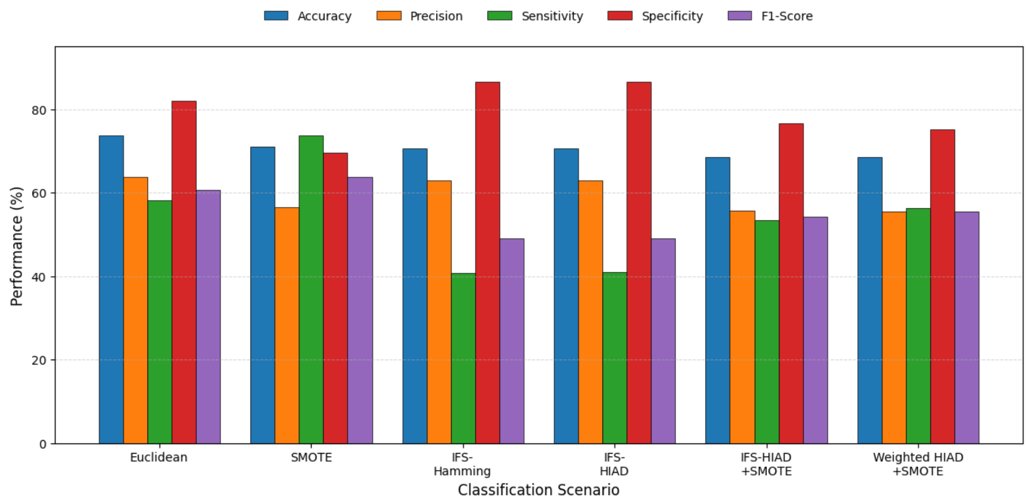

Fig. 3: Average Accuracy, Precision, Sensitivity, Specificity, and F1-Score across all evaluated neighborhood configurations ($K \in \{3, 5, 7, 9\}$). Reported values correspond to the arithmetic mean obtained from the Stratified 5-Fold Cross-Validation experiments.

As illustrated in Fig. 3, the evaluated classification scenarios exhibit distinct performance profiles across the five evaluation metrics. The conventional K-NN (Euclidean) configuration achieved the highest overall accuracy, whereas the K-NN + SMOTE model produced the highest sensitivity and F1-Score. In contrast, the IFS-based configurations employing Hamming

Table 3: Performance comparison of all classification scenarios using Stratified 5-Fold Cross-Validation (%)

K	Scenario	Accuracy	Precision	Sensitivity	Specificity	F1-Score
3	K-NN (Euclidean)	73.57 ± 2.65	62.72 ± 4.00	60.07 ± 7.32	80.80 ± 3.06	61.16 ± 4.76
	K-NN + SMOTE	69.91 ± 3.33	55.59 ± 4.07	68.64 ± 7.49	70.60 ± 3.50	61.31 ± 4.86
	IFS-Hamming	67.71 ± 2.10	55.53 ± 5.15	42.56 ± 6.12	81.20 ± 5.08	47.72 ± 3.55
	IFS-HIAD	67.71 ± 2.10	55.53 ± 5.15	42.56 ± 6.12	81.20 ± 5.08	47.72 ± 3.55
	IFS-HIAD + SMOTE	65.88 ± 1.89	51.45 ± 2.99	54.15 ± 6.41	72.20 ± 5.64	52.41 ± 2.24
	Weighted HIAD + SMOTE	65.23 ± 2.61	50.67 ± 3.38	56.38 ± 6.27	70.00 ± 6.96	52.99 ± 1.90
5	K-NN (Euclidean)	73.57 ± 2.43	63.02 ± 4.12	59.34 ± 2.50	81.20 ± 2.93	61.08 ± 2.95
	K-NN + SMOTE	70.57 ± 2.99	56.23 ± 4.08	73.51 ± 5.06	69.00 ± 5.06	63.56 ± 3.04
	IFS-Hamming	70.70 ± 3.06	62.62 ± 7.23	41.05 ± 4.29	86.60 ± 3.72	49.44 ± 4.68
	IFS-HIAD	70.70 ± 3.06	62.62 ± 7.23	41.05 ± 4.29	86.60 ± 3.72	49.44 ± 4.68
	IFS-HIAD + SMOTE	68.88 ± 2.35	55.59 ± 3.09	53.78 ± 7.37	77.00 ± 3.16	54.46 ± 4.78
	Weighted HIAD + SMOTE	68.62 ± 2.32	54.85 ± 3.33	59.04 ± 9.03	73.80 ± 4.66	56.49 ± 4.55
7	K-NN (Euclidean)	73.82 ± 2.67	64.38 ± 5.04	57.07 ± 4.88	82.80 ± 4.21	60.31 ± 3.78
	K-NN + SMOTE	72.13 ± 3.29	57.88 ± 4.14	75.00 ± 6.62	70.60 ± 4.27	65.21 ± 4.20
	IFS-Hamming	72.26 ± 3.42	66.75 ± 7.89	41.82 ± 7.04	88.60 ± 3.72	51.09 ± 6.42
	IFS-HIAD	72.39 ± 3.59	66.89 ± 7.97	42.19 ± 7.74	88.60 ± 3.72	51.38 ± 6.91
	IFS-HIAD + SMOTE	69.27 ± 1.55	56.55 ± 2.28	52.68 ± 7.21	78.20 ± 3.92	54.25 ± 3.79
	Weighted HIAD + SMOTE	69.79 ± 1.74	57.10 ± 2.66	54.91 ± 7.36	77.80 ± 3.76	55.70 ± 4.03
9	K-NN (Euclidean)	73.96 ± 2.37	64.99 ± 5.26	56.37 ± 3.56	83.40 ± 4.08	60.19 ± 2.74
	K-NN + SMOTE	71.48 ± 3.42	56.80 ± 4.02	77.61 ± 6.95	68.20 ± 4.53	65.46 ± 4.18
	IFS-Hamming	71.61 ± 2.60	66.52 ± 6.96	37.69 ± 5.10	89.80 ± 2.32	48.00 ± 5.42
	IFS-HIAD	71.61 ± 2.60	66.52 ± 6.96	37.69 ± 5.10	89.80 ± 2.32	48.00 ± 5.42
	IFS-HIAD + SMOTE	70.31 ± 3.23	58.91 ± 5.46	53.40 ± 5.54	79.40 ± 6.15	55.63 ± 3.24
	Weighted HIAD + SMOTE	70.70 ± 1.81	59.03 ± 3.76	54.90 ± 5.70	79.20 ± 4.79	56.56 ± 2.37

and HIAD distances achieved the highest specificity values. Meanwhile, the Weighted HIAD + SMOTE configuration exhibited a more balanced sensitivity–specificity profile, suggesting an intermediate classification behavior between minority-class detection and healthy-patient identification.

The results presented in Table 3 indicate that the conventional K-NN (Euclidean) classifier achieved the highest overall accuracy, reaching 73.96% at $K = 9$. However, despite its relatively strong overall performance, the model exhibited only moderate sensitivity values ranging from 56.37% to 60.07%, indicating a limited ability to identify diabetic patients from the minority class.

The incorporation of SMOTE substantially improved minority-class recognition. Across all neighborhood configurations, the K-NN + SMOTE model consistently achieved the highest sensitivity values, reaching 77.61% at $K = 9$, while also obtaining the highest F1-Score of 65.46%. These findings suggest that synthetic minority oversampling can alleviate the effects of class imbalance and improve the detection of diabetic cases. Nevertheless, this improvement was accompanied by lower specificity values, indicating a greater tendency to classify patients as diabetic.

A different classification behavior was observed for the IFS-based configurations employing Hamming and HIAD distances. Both approaches consistently achieved the highest specificity values, reaching 89.80% at $K = 9$. This result suggests that the intuitionistic fuzzy representation tends to produce a more conservative classification tendency that reduces false-positive predictions. However, such conservatism was also associated with lower sensitivity values, indicating that a considerable proportion of diabetic patients remained undetected.

Interestingly, the IFS-Hamming and IFS-HIAD configurations produced highly comparable classification results across all neighborhood configurations. At several values of K, both approaches yielded identical values for Accuracy, Precision, Sensitivity, Specificity, and F1-Score. This similarity suggests that both formulations induce highly comparable neighborhood structures under the current IFS representation. Therefore, within the experimental setting considered in this study, HIAD should be interpreted as a Hausdorff-inspired alternative similarity formulation rather than a distance measure that consistently outperforms the conventional Hamming distance.

The Weighted HIAD + SMOTE configuration produced a more balanced trade-off between sensitivity and specificity than the unweighted IFS-based models. For example, at $K = 9$, sensitivity increased from 37.69% under IFS-HIAD to 54.90% under Weighted HIAD + SMOTE, while specificity decreased from 89.80% to 79.20%. These results suggest that incorporating clinically informed feature weighting can improve diabetic patient detection while maintaining relatively high specificity. However, the observed differences should be interpreted as a redistribution of the sensitivity–specificity trade-off rather than as a clear overall performance advantage over alternative distance formulations.

Overall, the experimental findings indicate that different methodological components contribute different classification tendencies. In the evaluated setting, SMOTE was associated with improved detection of diabetic cases, whereas the intuitionistic fuzzy representation was associated with stronger identification of healthy patients. Meanwhile, the incorporation of clinically informed feature weighting contributed to a more balanced sensitivity–specificity profile. These observations highlight the importance of considering multiple evaluation metrics when assessing uncertainty-aware medical classification frameworks.

3.2. Sensitivity–Specificity Trade-off Analysis

In medical screening applications, overall accuracy alone may not adequately reflect diagnostic performance because classifiers must balance diabetic patient detection and healthy-patient identification. Consequently, sensitivity and specificity were analyzed jointly to examine the classification tendencies of the evaluated scenarios.

Fig. 4 presents the sensitivity and specificity obtained across different neighborhood configurations. The results indicate that different methodological components were associated with distinct sensitivity–specificity profiles.

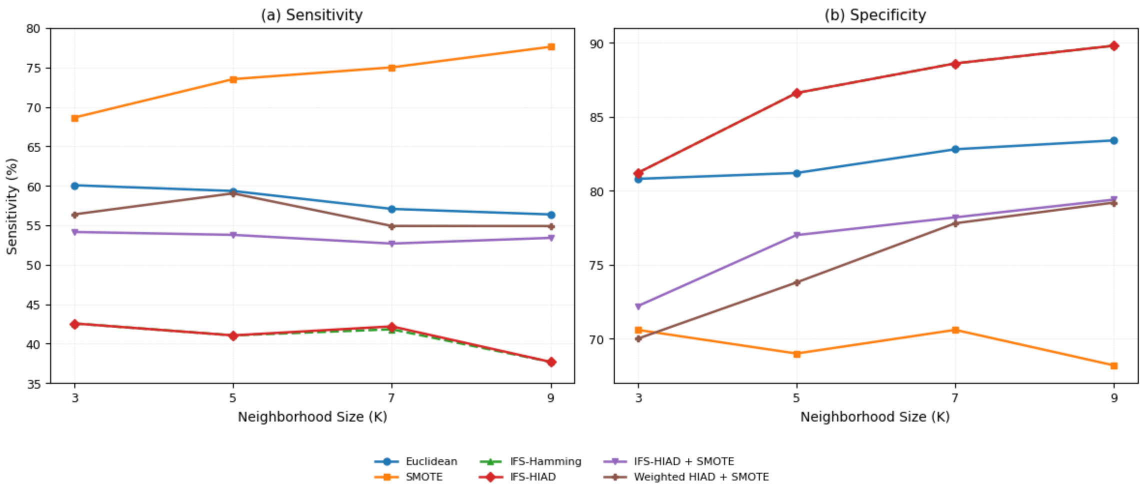

Fig. 4: Influence of neighborhood size ($K \in \{3, 5, 7, 9\}$) on (a) sensitivity and (b) specificity across the six evaluated classification scenarios.

As shown in Fig. 4(a), the incorporation of SMOTE consistently increased sensitivity across all neighborhood configurations, indicating improved detection of diabetic patients. However, this improvement was accompanied by lower specificity values, as illustrated in Fig. 4(b), reflecting a greater tendency to classify patients as diabetic. This behavior is consistent with the objective of synthetic minority oversampling, which aims to improve minority-class recognition by reducing the imbalance between diabetic and non-diabetic observations.

In contrast, the IFS-based configurations employing Hamming and HIAD distances achieved higher specificity values than the Euclidean-based models. The substantial overlap between the Hamming-based and HIAD-based curves further indicates that both formulations produced highly similar classification behavior across all evaluated neighborhood configurations. Although this conservative tendency reduced false-positive predictions, it was also associated with lower sensitivity, indicating that a larger proportion of diabetic patients remained undetected compared with the SMOTE-enhanced configurations.

The Weighted HIAD + SMOTE configuration exhibited an intermediate sensitivity–specificity profile between the highly sensitive SMOTE-based model and the highly specific unweighted IFS-based models. While the numerical differences relative to IFS-HIAD + SMOTE were modest, the weighted formulation demonstrated that incorporating clinically informed feature importance may influence the distribution of classification decisions within the proposed framework. This observation suggests that feature weighting can partially adjust the balance between minority-class detection and healthy-patient identification while preserving the uncertainty-aware characteristics of the IFS representation.

Overall, the results indicate that different methodological components contribute distinct classification tendencies. In the evaluated setting, SMOTE was primarily associated with improved detection of diabetic patients, whereas the intuitionistic fuzzy representation was associated with stronger identification of healthy patients. The Weighted HIAD + SMOTE configuration provided an intermediate sensitivity–specificity profile, suggesting that clinically informed feature weighting can partially moderate the trade-off between minority-class detection and false-positive reduction. These findings highlight the importance of jointly considering sensitivity and specificity when evaluating uncertainty-aware medical classification frameworks.

3.3. Comparative Analysis of Hamming Distance and HIAD

The empirical evaluation of Scenarios 3 (IFS-Hamming) and 4 (IFS-HIAD) provides insight into the relationship between the conventional Hamming distance and the proposed Hausdorff-Inspired Attribute Distance (HIAD) within the Intuitionistic Fuzzy Set (IFS) framework. As shown in Table 3, both approaches produced highly comparable classification results across all evaluated neighborhood configurations. At $K = 3$, $K = 5$, and $K = 9$, the two methods yielded identical performance across all reported evaluation metrics, whereas only marginal differences were observed at $K = 7$. For example, the HIAD-based configuration achieved an accuracy of 72.39% compared with 72.26% for the Hamming-based configuration. Consistent with these numerical results, the substantial overlap between the Hamming-based and HIAD-based curves in Fig. 4 further indicates highly similar classification behavior.

This similarity can be explained by the structure of the proposed IFS representation. Both Hamming distance and HIAD operate on the same intuitionistic components, namely membership (μ), non-membership (ν), and hesitation (π). Because the non-membership and hesitation components are derived directly from the membership function through the control parameter λ, all three components remain strongly dependent on the same underlying information. Consequently, both distance formulations produce highly similar pairwise similarity relationships. Since K-NN classification depends primarily on neighborhood ranking rather than absolute distance magnitudes, similar pairwise similarities naturally induce similar neighborhood structures and, therefore, comparable classification outcomes.

Conceptually, however, HIAD differs from the conventional Hamming distance. While Hamming distance aggregates absolute deviations across intuitionistic components, HIAD emphasizes the maximum deviation among the membership, non-membership, and hesitation values within each attribute. By preserving the largest intuitionistic discrepancy before global aggregation, HIAD provides an alternative uncertainty-aware similarity formulation inspired by the maximum-deviation principle of the Hausdorff distance. Therefore, the contribution of HIAD should be interpreted as providing an alternative similarity-computation mechanism within the IFS framework rather than as a formulation intended to consistently outperform conventional IFS-based distance measures.

3.4. Impact of Neighborhood Size (K)

The experimental results indicate that neighborhood size (K) influences classification behavior across all evaluated scenarios. Since K-NN assigns class labels based on neighboring instances, varying K affects the balance between local sensitivity and neighborhood stability, leading to different classification tendencies across data representations and similarity formulations.

Fig. 4 illustrates the sensitivity and specificity obtained across different neighborhood configurations. While some scenarios exhibited relatively stable performance, others showed noticeable changes as the neighborhood size increased.

For the conventional Euclidean-based K-NN classifier (Scenario 1), increasing K was associated with a slight improvement in accuracy, from 73.57% at $K = 3$ to 73.96% at $K = 9$, while sensitivity and specificity remained relatively stable. This result suggests that larger neighborhoods provide more stable majority-voting decisions without substantially altering classification behavior.

A more pronounced effect was observed in the SMOTE-enhanced configuration (Scenario 2). Sensitivity increased steadily from 68.64% at $K = 3$ to 77.61% at $K = 9$, whereas specificity decreased slightly from 70.60% to 68.20%. These findings indicate that larger neighborhood sizes strengthened minority-class recognition when class balancing was applied, although at the cost of reduced healthy-patient identification.

In contrast, the IFS-based configurations employing Hamming and HIAD distances (Scenarios 3 and 4) exhibited increasing specificity as K increased, reaching 89.80% at $K = 9$, while sensitivity gradually declined. The highly similar performance trends observed for both formulations are consistent with the findings of the previous subsection, suggesting that the two distance measures generated comparable neighborhood structures under the proposed IFS representation.

The introduction of SMOTE into the HIAD-based IFS framework (Scenario 5) increased sensitivity relative to the unbalanced IFS configuration while moderately reducing specificity, indicating improved minority-class recognition while preserving the overall classification characteristics of the IFS representation.

The Weighted HIAD + SMOTE configuration (Scenario 6) demonstrated comparatively stable performance across different neighborhood sizes. Accuracy increased from 65.23% at $K = 3$ to 70.70% at $K = 9$, while sensitivity remained within a relatively narrow range, suggesting that clinically informed feature weighting contributed to a more consistent sensitivity–specificity profile.

Overall, the results demonstrate that neighborhood size constitutes an important hyperparameter within both conventional and intuitionistic fuzzy K-NN frameworks. Larger neighborhood sizes favored sensitivity in the SMOTE-enhanced configuration, whereas they favored specificity in the IFS-based models. Consequently, no single value of K consistently produced the best performance across all evaluation metrics, and its selection should therefore be guided by the intended diagnostic objective.

3.5. Clinical Implications and Study Limitations

The experimental findings highlight the importance of balancing diabetic patient detection and healthy-patient identification in medical screening systems. Because classification models often operate under uncertainty arising from overlapping patient characteristics, incomplete clinical information, and class imbalance, overall accuracy alone may not adequately reflect diagnostic performance. In particular, false-negative predictions may delay diagnosis and treatment, whereas excessive false-positive predictions may increase healthcare costs and unnecessary follow-up examinations. Therefore, sensitivity and specificity should be considered jointly when evaluating medical classification systems.

The results further demonstrate the potential of Intuitionistic Fuzzy Set (IFS) representations for handling uncertainty in diabetes classification. By simultaneously modeling membership, non-membership, and hesitation components, the IFS framework provides an additional mechanism for representing ambiguous clinical conditions. Within this framework, the proposed Hausdorff-Inspired Attribute Distance (HIAD) offers an alternative maximum-deviation-based similarity formulation. Although its performance was highly comparable to that of the conventional Hamming distance, HIAD expands the range of similarity measures available within the IFS framework while preserving comparable classification performance.

To evaluate parameter robustness, a localized sensitivity analysis was conducted using $\lambda \in \{0.5, 0.6, 0.7, 0.8\}$ within the Weighted HIAD + SMOTE configuration. As shown in Eq. (4), identical values of accuracy, precision, sensitivity, specificity, and F1-score were obtained across all evaluated values of λ. This result suggests that moderate redistributions between the non-membership and hesitation components did not substantially alter the induced neighborhood structure within the investigated range, leaving the classification outcomes unchanged.

Table 4: Localized sensitivity analysis of the non-membership control parameter (λ) for the proposed Weighted HIAD + SMOTE configuration at $K = 9$.

λ	Accuracy	Precision	Sensitivity	Specificity	F1-Score
0.5	70.70	59.03	54.90	79.20	56.56
0.6	70.70	59.03	54.90	79.20	56.56
0.7	70.70	59.03	54.90	79.20	56.56
0.8	70.70	59.03	54.90	79.20	56.56

Several limitations should be acknowledged. First, all experiments were conducted using the Pima Indians Diabetes dataset, limiting the generalizability of the findings to other populations and healthcare environments. Broader validation on additional medical datasets is therefore required before stronger conclusions regarding practical clinical applicability can be drawn. Second, the feature-weighting strategy was based on clinically informed heuristic assumptions rather than data-driven optimization procedures. Third, the sensitivity analysis was restricted to a relatively narrow range of λ values. Finally, the proposed HIAD formulation was evaluated only against the conventional Hamming distance within the IFS framework.

Future research may extend the proposed framework by incorporating adaptive feature-weighting mechanisms, larger and more diverse medical datasets, alternative neighborhood optimization strategies, and broader comparisons with uncertainty-aware similarity measures and classification methods.

4. Conclusion

This study investigated the application of Intuitionistic Fuzzy Set (IFS) representations and the proposed Hausdorff-Inspired Attribute Distance (HIAD) within a K-Nearest Neighbor (K-NN) framework for diabetes classification. Experimental evaluations were conducted on the Pima Indians Diabetes dataset using six classification scenarios and multiple neighborhood configurations ($K \in \{3, 5, 7, 9\}$) under Stratified 5-Fold Cross-Validation.

The results demonstrated that different methodological components contributed distinct classification tendencies. The conventional Euclidean-based K-NN configuration achieved the highest overall accuracy, whereas incorporating SMOTE improved sensitivity in detecting diabetic patients. In contrast, the IFS-based configurations employing Hamming and HIAD distances achieved higher specificity values, indicating stronger healthy-patient identification under the evaluated conditions.

The comparative analysis showed that the proposed HIAD formulation produced classification performance highly comparable to that of the conventional Hamming distance across all evaluated neighborhood configurations. These findings suggest that HIAD serves as an alternative maximum-deviation-based similarity formulation within the IFS framework, rather than one intended to consistently outperform existing distance measures. Furthermore, the integration of class balancing and clinically informed feature weighting resulted in a more balanced sensitivity–specificity profile than the unweighted IFS-based configurations.

A localized sensitivity analysis demonstrated identical classification performance across the evaluated values of the non-membership control parameter (λ), suggesting that moderate redistributions between the non-membership and hesitation components did not substantially alter the induced neighborhood structure within the investigated range.

Despite these encouraging findings, the present study was limited to a single benchmark dataset and a heuristic feature-weighting strategy. Future research may focus on validating the proposed framework with larger, more diverse medical datasets, exploring adaptive feature weighting approaches, and conducting broader comparisons with uncertainty-aware similarity measures and classification methods.

Overall, the proposed HIAD formulation expands the range of similarity measures available within the intuitionistic fuzzy framework and provides an alternative approach for uncertainty-aware diabetes classification.

CRediT Authorship Contribution Statement

Tio Valent Novi Yanti Sianturi: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing–Original Draft, Visualization. **Raden Sulaiman:** Validation, Supervision, Project Administration, Review & Editing.

Declaration of Generative AI and AI-assisted Technologies

During the preparation of this manuscript, the authors used Generative AI tools (ChatGPT by OpenAI and Gemini by Google) to enhance language quality, including grammar correction, paraphrasing, and improving clarity and readability. The use of these tools was strictly limited to linguistic refinement. All core research components, including research design, data processing, statistical analysis, interpretation of results, and scientific conclusions, were developed independently by the authors. The authors remain fully responsible for the originality, accuracy, and integrity of the final manuscript.

Declaration of Competing Interest

The authors declare no competing interests.

Funding and Acknowledgments

This research received no external funding.

Data and Code Availability

The dataset used in this study is publicly available through the UCI Machine Learning Repository and Kaggle. The source code developed for the implementation of the Intuitionistic Fuzzy Set (IFS) framework, the Hausdorff-Inspired Attribute Distance (HIAD), the Weighted HIAD formulation, and the experimental evaluation pipeline is available from the corresponding author upon reasonable request.

References

[1] World Health Organization. *Diabetes: Key Facts*. https://www.who.int/news-room/fact-sheets/detail/diabetes. Accessed: 2024-05-10. 2024.

[2] International Diabetes Federation. *IDF diabetes atlas, 10th edition*. https://diabetesatlas.org. Accessed: 2026-04-14. 2021.

[3] Xueyan Wang, Ping Shen, Guoxu Zhao, et al. "An enhanced machine learning algorithm for type 2 diabetes prognosis with a detailed examination of key correlates". In: *Scientific Reports* 14.1 (2024), p. 26355. DOI: 10.1038/s41598-024-75898-w.

[4] American Diabetes Association. "Classification and diagnosis of diabetes: Standards of medical care in diabetes—2022". In: *Diabetes Care* 45.Supplement_1 (2022), S17–S38. DOI: 10.2337/dc22-S002.

[5] Jelena Tasic, Zsófia Nagy-Perjési, and Márta Takács. "Multilevel Fuzzy Inference System for Estimating Risk of Type 2 Diabetes". In: *Mathematics* 12.8 (2024), p. 1167. DOI: 10.3390/math12081167.

[6] Saurabh Suman, Utkarsha Pacharaney, and Pradnyawant M. Gote. "A type-1 fuzzy logic framework for early risk assessment of gestational diabetes mellitus using clinical risk factors". In: *Proceedings of the 2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS)*. Coimbatore, India: IEEE, 2025. DOI: 10.1109/ICMLAS64557.2025.10967700.

[7] Haiyan Zhao, Qian Xiao, Zheng Liu, and Yanhong Wang. "An approach in medical diagnosis based on Z-numbers soft set". In: *PLOS ONE* 17.8 (2022), e0272203. DOI: 10.1371/journal.pone.0272203.

[8] M. Kamrul Hasan, Md. Ashraful Alam, Dola Das, Eklas Hossain, and Mahmudul Hasan. "Diabetes prediction using ensembling of different machine learning classifiers". In: *IEEE Access* 8 (2020), pp. 76516–76531. DOI: 10.1109/ACCESS.2020.2989857.

[9] Xin Feng, Yihuai Cai, and Ruihao Xin. "Optimizing diabetes classification with a machine learning-based framework". In: *BMC Bioinformatics* 24.1 (Oct. 2023), p. 428. DOI: 10.1186/s12859-023-05467-x.

[10] Michel Marie Deza and Elena Deza. *Encyclopedia of Distances*. 4th. Berlin, Heidelberg: Springer, 2016. DOI: 10.1007/978-3-662-52844-0.

[11] Eulalia Szmidt. *Distances and Similarities in Intuitionistic Fuzzy Sets*. Vol. 307. Studies in Fuzziness and Soft Computing. Cham, Switzerland: Springer, 2014. DOI: 10.1007/978-3-319-01640-5.

[12] Kemal Polat and Salih Güneş. "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease". In: *Digital Signal Processing* 17.4 (2007), pp. 702–710. DOI: 10.1016/j.dsp.2006.09.005.

[13] Nisha Gupta, Harjeet Singh, and Jimmy Singla. "Fuzzy Logic-based Systems for Medical Diagnosis – A Review". In: *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*. Coimbatore, India: IEEE, 2022. DOI: 10.1109/ICESC54411.2022.9885338.

[14] Krassimir T. Atanassov. "Intuitionistic fuzzy sets". In: *Fuzzy Sets and Systems* 20.1 (1986), pp. 87–96. DOI: 10.1016/S0165-0114(86)80034-3.

[15] T.-Y. Chen. "A circular intuitionistic fuzzy evaluation method based on distances from the average solution to support multiple criteria intelligent decisions involving uncertainty". In: *Engineering Applications of Artificial Intelligence* 117 (2023), p. 105499. DOI: 10.1016/j.engappai.2022.105499.

[16] Huiping Chen and Yan Liu. "Group decision-making method with incomplete intuitionistic fuzzy soft information for medical diagnosis model". In: *Mathematics* 12.12 (2024), p. 1823. DOI: 10.3390/math12121823.

[17] Przemysław Grzegorzewski. "Distances between intuitionistic fuzzy sets and/or interval-valued fuzzy sets based on the Hausdorff metric". In: *Fuzzy Sets and Systems* 148.2 (2004), pp. 319–328. DOI: 10.1016/j.fss.2003.08.005.

[18] Xinxing Wu, Zhiyi Zhu, and Shyi-Ming Chen. "Strict intuitionistic fuzzy distance similarity measures based on Jensen-Shannon divergence". In: *Information Sciences* 661 (2024), p. 120144. DOI: 10.1016/j.ins.2024.120144.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357. DOI: 10.1613/jair.953.

[20] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes. *Pima Indians Diabetes*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/52/pima+indians+diabetes. Accessed: 2026-04-14. 1988.

[21] Kaggle. *Pima Indians Diabetes Database*. Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Accessed: 2026-04-14. 2016.

[22] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. 3rd. The Morgan Kaufmann Series in Data Management Systems. Waltham, MA: Morgan Kaufmann, 2012. DOI: 10.1016/C2009-0-61819-5.

[23] Xinxing Wu, Huan Tang, Zhiyi Zhu, Lantian Liu, Guanrong Chen, and Miin-Shen Yang. "Nonlinear strict distance and similarity measures for intuitionistic fuzzy sets with applications to pattern classification and medical diagnosis". In: *Scientific Reports* 13 (2023), p. 13918. DOI: 10.1038/s41598-023-40817-y.

[24] Hans-Jürgen Zimmermann. *Fuzzy Set Theory and Its Applications*. Springer, 2010. DOI: 10.1007/978-94-010-0646-0.