

# Enhanced Missing Data Imputation Using Intuitionistic Fuzzy Rough-Nearest Neighbor Approach

Shivani Singh <sup>a</sup>

AN & SK School of Information Technology, Indian Institute of Technology Delhi, New Delhi, India

**Keywords:** Intuitionistic Fuzzy Rough Sets, k-Nearest Neighbourhood, Missing Data Imputation.


**Abstract:** The exponential growth of databases across various domains necessitates robust techniques for handling missing data to maintain data integrity and analytical accuracy. Traditional approaches often struggle with real-valued datasets due to inherent limitations in handling uncertainty and imprecision. Nearest Neighbourhood algorithms have proven beneficial in missing data imputation, offering effective solutions to address data gaps. In this paper, we propose a novel method for missing data imputation, termed Intuitionistic Fuzzy Rough-Nearest Neighbourhood Imputation (IFR-NNI), which extends the application of intuitionistic fuzzy rough sets to handle missing data scenarios. By integrating Intuitionistic Fuzzy Rough Sets into the nearest neighbor imputation framework, we aim to overcome the limitations of traditional methods, including information loss, challenges in managing uncertainty and vagueness, and instability in approximation outcomes. The proposed method is implemented on real-valued datasets, and non-parametric statistical analysis is performed to evaluate its performance. Our findings indicate that the IFR-NNI method demonstrates excellent performance in general, showcasing its effectiveness in addressing missing data scenarios and advancing the field of data imputation methodologies.

## 1 INTRODUCTION

The extraction of meaningful insights from data is fundamental for understanding phenomena and facilitating processes such as classification and regression. Across diverse domains including science, communication, and business, vast amounts of data are generated and utilized. However, datasets frequently encounter missing data due to various factors such as input errors, faulty measurements, or non-responses in assessments. For instance, in wireless sensor networks, missing data is often inevitable due to sensor faults or communication malfunctions (Li and Parker, 2014), while in DNA microarray studies, missing data may arise from insufficient resolution or image corruption (Sun et al., 2010). Additionally, repositories like the UCI Machine Learning Repository commonly contain datasets with substantial proportions of missing values. The presence of missing values poses significant challenges, particularly in the context of machine learning techniques, where interpretation and analysis may be severely compromised. Consequently, missing data imputation emerges as a critical issue across

scientific research communities, particularly in data mining and machine learning domains (Aydilek and Arslan, 2012; Nelwamondo et al., 2013).

Addressing missing values can be approached in various ways. While simple strategies like deletion or substitution with zero or mean values are common, they often lead to information loss and bias in assessments. Alternatively, imputation methods aim to estimate missing values using statistical or machine learning approaches. The nature of missing data can be categorized into three types: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) (Little and Rudin, 2019). Understanding these categories is crucial for selecting appropriate imputation techniques. Statistical methods typically employ simple approaches like mean or mode imputation, while machine learning-based methods involve building models to predict missing values (García-Laencina et al., 2010). Nearest Neighbour (NN) based methods have gained popularity for missing value imputation due to their accuracy and simplicity. However, they require specifying the number of neighbors and suffer from high time

<sup>a</sup>  <https://orcid.org/0000-0001-7054-1193>

complexity and local optima issues. Conversely, statistical methods may introduce bias and complexity, relying on initial guesses and eigenvector representations (Troyanskaya et al., 2001).

The use of rough set theory introduced by Pawlak (2012) for missing data imputation is motivated by its strength in handling vagueness and incompleteness in data without requiring additional information. It provides robust approximations and decision rules directly from the dataset, ensuring both effectiveness and interpretability. The use of intuitionistic rough sets, rather than classical rough sets, further enhances this capability by addressing both uncertainty and vagueness through the inclusion of membership and non-membership functions. This dual aspect offers a more nuanced approximation, particularly useful in scenarios with incomplete or imprecise data, where classical rough sets may not fully capture the inherent uncertainty.

In this paper, we introduce a novel approach to missing data imputation, leveraging the combination of Intuitionistic Fuzzy (IF) rough sets and the nearest neighbour algorithm. By integrating IF rough sets with NN estimation, we aim to capitalize on the accuracy of NN methods while enhancing noise tolerance and robustness. Specifically, we propose IF rough-nearest neighbour imputation methods. The subsequent sections of this paper are organized as follows: Section 2 reviews relevant literature. Section 3 provides essential preliminaries to understand the theoretical background. Section 4 introduces the proposed methodologies. Section 5 presents the implementation of these methods on benchmark datasets and evaluates their performance using non-parametric statistical analysis. Finally, Section 6 concludes our work and outlines future research directions.

## 2 LITERATURE REVIEW

Various domains such as meteorology, transportation, and others have witnessed the treatment of missing-valued data by researchers. Although several algorithms with different approaches have been proposed, they are not commonly employed for specific domains or datasets. Notable imputation techniques frequently used across fields include those based on Nearest Neighbours (NN), which predict missing values based on neighboring instances. While NN methods offer accuracy and simplicity, they come with drawbacks such as the need for specifying the number of

neighbors, high time complexity, and local optima issues.

Troyanskaya et al. (2001) proposed two methods, KNN and SVD, for imputation in DNA microarrays. KNN computes a weighted average of values based on Euclidean distance from the  $K$  closest genes, while SVD employs an expectation maximization (EM) algorithm to approximate missing values. Comparing the two, KNN showed greater robustness, particularly with increasing percentages of missing values. Batista and Monard (2003) introduced the  $k$ -nearest neighbor imputation (KNNI) method, which replaces missing values with the mean value of specific attribute neighbors. Grzymala-Busse (2005) introduced global most common (GMC), global most common average (GMCA) methods for nominal and numeric attributes, respectively, where missing values are replaced by the most common or average attribute values. Kim et al. (2005) proposed the local least squares imputation (LLSI) method, which estimates missing attribute values as a linear combination of similar genes selected through  $k$ -nearest neighbors.

Schneider (2001) introduced an algorithm based on regularized Expectation-Maximization (EM) for missing value prediction, utilizing Gaussian distribution to parameterize data and iteratively maximizing likelihoods. Oba et al. (2003) proposed Bayesian PCA imputation (BPCAI), incorporating Bayesian estimation into the approximation stage. Honghai et al. (2005) presented SVM-based imputation methods, utilizing Support Vector Machines and Support Vector Regressors. Clustering-based methods, such as those by Li et al. (2004) and Liao et al. (2009), use techniques like  $K$ -means and Fuzzy  $k$ -means for imputation, often incorporating sliding window mechanisms for data stream handling. Neural network-based methods, including Multi-Layer Perceptrons (MLP) (Sharpe and Sholly, 1995), Recurrent Neural Networks (RNN) (Bengio and Gingras, 1995), and Auto Associative Neural Networks (AANN) (Pyle, 1999), have been employed for imputation, each with its own approach and advantages. Amiri and Jensen (2016) introduced fuzzy rough set-based nearest neighbor algorithms for imputation, showing superior performance compared to traditional methods. In the paper (Pereira et al., 2020), the adaptability of Autoencoders in handling various types of missing data are discussed.

While clustering-based algorithms often exhibit high computational complexity, those based on nearest neighbors are preferred for their computational efficiency. Intuitionistic Fuzzy (IF) set theory, known for effectively handling vagueness and

uncertainty, remains unexplored in missing value imputation. In this work, we propose a missing data imputation method based on IF rough-nearest neighbor approach.

### 3 PRELIMINARIES

In this section a basic overview on Intuitionistic fuzzy rough sets (IFRS) is given.

**Definition 3.1.** (Huang, 2013): A quadruple  $IS = (U, AT, V, h)$  is called an Information System, where  $U = \{u_1, u_2, \dots, u_n\}$  is a non-empty finite set of objects, called the universe of discourse,  $AT = \{a_1, a_2, \dots, a_m\}$  is a non-empty finite set of attributes.  $V = \bigcup_{a \in AT} V_a$  where  $V_a$  is the set of attribute values associated with each attribute  $a \in AT$  and  $h: U \times AT \rightarrow V$  is an information function that assigns particular values to the objects against attribute set such that  $\forall a \in AT, \forall u \in U$  and  $h(u, a) \in V_a$ .

An information system  $IS = (U, AT, V, h)$  is said to be a Decision System if  $AT = C \cup D$  where  $C$  is a non-empty finite set of conditional features/attributes and  $D$  is a non-empty collection of decision features/attributes with  $C \cap D = \emptyset$ . Here  $V = V_C \cup V_D$  with  $V_C$  and  $V_D$  as the set of conditional attribute values and decision attribute values, respectively.

**Definition 3.2.** (Pawlak, 2012): Let  $IS = (U, AT, V, h)$  be a decision system. For  $P \subset AT$ , a P-indiscernibility relation is defined as:

$$R_P = (x, y) \in U * U | \forall p \in P \Rightarrow p(x) = p(y)$$

where,  $R_P$  is an equivalence relation and  $[x]_{R_P}$  divides the set  $U$  into equivalence classes defined by the attributes belongs to  $P$ . If  $A \subseteq U$ , then the lower and upper approximation of set  $A$  are given by:

$$\begin{aligned} \underline{(R_P)A} &= \{x \in U | [x]_{R_P} \subseteq A\} \\ \overline{(R_P)A} &= \{x \in U | [x]_{R_P} \cap A \neq \emptyset\} \end{aligned}$$

All the data instances that contained in set  $\underline{(R_P)A}$  must contained in set  $A$  while the instances that contained in  $\overline{(R_P)A}$  may be a member of  $A$ .

**Definition 3.3.** (Atanassov, 1999): Given a non-empty finite universe of discourse  $U$ . A set  $A$  on  $U$  having the form  $A = \{(x, \mu_A(x), \nu_A(x)) | x \in U\}$  is said to be an IF set, where  $\mu_A : U \rightarrow [0, 1]$  and  $\nu_A : U \rightarrow [0, 1]$  with the condition  $0 \leq \mu_A(x) + \nu_A(x) \leq$

$1, \forall x \in U$  are known as membership degree and non-membership degree of the element  $x$  in  $A$ , respectively.  $\pi_A(x) = 1 - \mu_A(x) - \nu_A(x)$  is the degree of hesitancy of the element  $x$  in IF set  $A$ .

The cardinality of an IF set  $A$  is given by  $|A| = \sum_{x \in A} \frac{1 + \mu_A(x) - \nu_A(x)}{2}$  where 1 in numerator is a translation factor that guarantees the positivity of  $|A|$  while 2 in denominator is a scaling factor which bounds the cardinality between 0 and 1.

An ordered pair  $\langle \mu, \nu \rangle$  is called an IF value, where  $0 \leq \mu + \nu \leq 1$  and  $0 \leq \mu, \nu \leq 1$ . An information system is said to be an IF information system if attribute values corresponding to objects are IF value.

**Properties:** For every two IF Sets  $A$  and  $B$  the following relations and operations hold:

1.  $A \subseteq B$  iff  $\mu_A(x) \leq \mu_B(x)$  and  $\nu_A(x) \geq \nu_B(x), \forall x \in U$
2.  $A = B$  iff  $A \subseteq B$  and  $B \subseteq A$
3.  $N(A) = \{(x, \nu_A(x), \mu_A(x)) | x \in U\}$
4.  $A + B = \{(x, \mu_A(x) + \mu_B(x) - \mu_A(x) \cdot \mu_B(x), \nu_A(x) \cdot \nu_B(x)) | x \in U\}$
5.  $A - B = \left\{ \left( \frac{\mu_A - \mu_B}{1 - \mu_B}, \frac{\nu_A}{\nu_B} \right), \text{if } \mu_A(x) \geq \mu_B(x), \nu_A(x) \leq \nu_B(x) > 0 \text{ and } \nu_A(x)\pi_B(x) \leq \pi_A(x)\nu_B(x) \right\}, \text{ and } , < 0, 1 > \text{ otherwise.}$
6.  $A \cdot B = \{(x, \mu_A(x) \cdot \mu_B(x), \nu_A(x) + \nu_B(x) - \nu_A(x) \cdot \nu_B(x)) | x \in U\}$
7.  $\lambda \cdot A = \{(x, 1 - (1 - \mu_A(x))^\lambda, (\nu_A(x))^\lambda) | x \in U\}$

where,  $N$  is a negation operator.

**Definition 3.4.** (Bustince and Burillo, 1996): An IF binary relation  $R(x_i, x_j) = \langle \mu_A(x_i, x_j), \nu_A(x_i, x_j) \rangle$  between objects  $x_i, x_j \in U$  is said to be an IF tolerance relation if it is reflexive (i.e.,  $\mu_A(x_i, x_i) = 1$  and  $\nu_A(x_i, x_i) = 0, \forall x_i \in X$ ) and symmetric (i.e.,  $\mu_A(x_i, x_j) = \mu_A(x_j, x_i)$  and  $\nu_A(x_i, x_j) = \nu_A(x_j, x_i), \forall x_i, x_j \in X$ ).

Let  $U$  be a collection of finite objects and  $C \subseteq A$ , an IF tolerance relation  $R_C(x_i, x_j) = \langle \mu_{R_C}(x_i, x_j), \nu_{R_C}(x_i, x_j) \rangle, c \in C$  is defined as:

$$\begin{aligned} \mu_{R_c}(x_i, x_j) &= \begin{cases} 1 - 2K(x_i, x_j), & K(x_i, x_j) \leq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \\ \nu_{R_c}(x_i, x_j) &= \begin{cases} 2((\mu_c(x_i) - \mu_c(x_j))^2 + (\nu_c(x_i) - \nu_c(x_j))^2), & K(x_i, x_j) \leq \frac{1}{2} \\ 1, & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

where,  $K(x_i, x_j) = (|\mu_c(x_i) - \mu_c(x_j)| + |\nu_c(x_i) - \nu_c(x_j)|)^2, \forall x_i, x_j \in U$  and  $R(x_i, x_j) = \min_{c \in A} R_c(x_i, x_j)$ .

**Definition 3.5.** (Cornelis et al., 2003): An IF triangular norm or IF t-norm  $T$  is a mapping from  $[0, 1] \times [0, 1] \rightarrow [0, 1]$  which is increasing, associative and commutative and satisfies  $T(1,x) = x, \forall x \in [0,1]$ .

An IF implicator  $I$  is a mapping  $[0, 1] \times [0, 1] \rightarrow [0, 1]$ , which is decreasing in its first component and increasing in second component with condition  $I(0, 0) = 1$  and  $I(1, x) = x, \forall x \in [0, 1]$ .

**Example 3.1:** If  $x = \langle x_1, x_2 \rangle$  and  $y = \langle y_1, y_2 \rangle$  in  $[0, 1]$  are two IF values then an IF t-norm and IF implicator are given as:

$$T(x, y) = [\max(0, x_1 + y_1 - 1), \min(1, x_2 + 1 - y_1, y_2 + 1 - x_1)] \tag{2}$$

$$I(x, y) = [\min(0, x_2 + y_1), \max(0, x_1 + y_2 - 1)] \tag{3}$$

**Definition 3.6.** (Cornelis et al., 2003): Given an IF set  $X \subseteq U$  and  $R(x_i, x_j)$  is an IF similarity/tolerance relation from  $U \times U \rightarrow [0,1]$  which assigns degree of similarity to each distinct pair of objects. The lower and upper approximation of  $X$  by  $R$  can be computed in many ways. A general definition is given as:

$$\underline{(R)}X(x_i) = \inf_{x_j \in U} \{I(R(x_i, x_j), X(x_j)), \forall x_i \in U\} \tag{4}$$

$$\overline{(R)}X(x_i) = \sup_{x_j \in U} \{I(R(x_i, x_j), X(x_j)), \forall x_i \in U\} \tag{5}$$

Here,  $I$  is an IF implicator and  $T$  is an IF t-norm and  $X(x_j) = 1$ , for  $x_j \in X$ , otherwise  $X(x_j) = 0$ . The pair  $\langle \underline{(R)}X(x_i), \overline{(R)}X(x_i) \rangle$  is called as IF rough set.

## 4 PROPOSED METHODOLOGY

Jensen and Cornelis introduced the model based on KNN algorithm using fuzzy-rough lower approximation and upper approximation in which discrete or continuous decision attribute values of datasets are predicted (Jensen and Cornelis, 2011). Based on this methodology Amiri and Jensen extended the FRNN model to predict the missing values presented in the dataset (Amiri and Jensen, 2016). We have further extended this KNN based algorithm for IF information system to impute the missing values in IF information system using IF rough sets approach.

In this subsection, IF rough approximation operators are defined to achieve the target of missing value imputation. This algorithm proposes that for each instance/object of the dataset consisting at least one missing value, that instance will be treated as

decision attribute and based on that attribute prediction is made. We address this algorithm as IF rough nearest neighbour imputation (IFRNNI).

### 4.1 If Rough Approximation Operators

**Definition 4.1:** The IF distance matrix  $d(y, z)$  for the difference between instances  $y \in U$  and  $z \in U$  in order to calculate the distance between the instances  $y = \langle \mu_1, \nu_1 \rangle$  and  $z = \langle \mu_n, \nu_n \rangle$  is defined as (Szmidski and Kacprzyk, 2001):

$$d(y, z) = \frac{\sqrt{(\mu_1 - \mu_n)^2 + (\nu_1 - \nu_n)^2 + (\pi_1 - \pi_n)^2 + equal(d(I_0), d(I_n))^2}}{\tag{6}}$$

where  $equal(d(I_0), d(I_n))^2 = \begin{cases} 0, & d(y) == d(z) \\ 1, & else \end{cases}$

**Definition 4.2:** IF similarity values for  $R(y, z)$ ,  $R_a(y, z)$ ,  $R_c(y, z)$  and  $R_d(y, z)$  with attribute  $a$ , conditional attributes  $c$ 's and decision attribute  $d$  are defined as:

$$R(y, z) = \min R_a(y, z) = \min (R_c(y, z), R_d(y, z)) \tag{7}$$

where  $R_a(y, z) = \begin{cases} < 1, 0 >, & d(y) == d(z) \\ < 0, 1 >, & else \end{cases}$

**Definition 4.3:** The lower approximation and upper approximation of instance  $y$  with respect to  $z$  are defined as in the following equations:

$$R \downarrow R_d z(y) = \inf_{p \in N} I(R(y, p), R_c(p, z)) \tag{8}$$

$$R \uparrow R_d z(y) = \sup_{p \in N} T(R(y, p), R_c(p, z)) \tag{9}$$

where,  $N$  is the  $k$ -nearest neighbour of instance  $y$ .  $R_d z$  is an IF tolerance relation which determines the similarities of two objects for the decision attribute.  $R_d(p, z)$  is also an IF tolerance relation which measures the similarity of objects  $z$  and  $p$  with respect to decision attribute  $d$ . In general,  $R_a z(p)$  signifies the similarity of objects  $z$  and  $p$  with respect to attribute  $a$ . Here, all IF tolerance relations are computed by Eq. (1).

One of the problems that are worth considering is in the process of computing the distance between two objects consisting of some missing attributes. Here, we simply avoid missing attributes while computing distances. Hence, the distance is only calculated between those instances having non-missing attribute values.

### 4.2 Prediction of Missing Values

**Definition 4.6.** With the help of lower and upper approximation operators,  $\tau_1$  and  $\tau_2$  are defined as

follows:

$$\begin{aligned}\tilde{\tau}_1 &= \langle \mu_{\tilde{\tau}_1}, \nu_{\tilde{\tau}_1} \rangle = \sum_{z \in N} \frac{[R \downarrow (R_d^z(y))] + [R \uparrow (R_d^z(y))]}{2} \times a(z) \\ \tilde{\tau}_2 &= \langle \mu_{\tilde{\tau}_2}, \nu_{\tilde{\tau}_2} \rangle = \sum_{z \in N} \frac{[R \downarrow (R_d^z(y))] + [R \uparrow (R_d^z(y))]}{2}\end{aligned}\quad (10)$$

**Definition 4.7.** The predicted missing value, namely  $\tilde{\tau}$ , obtained with the help of  $\tilde{\tau}_1$  and  $\tilde{\tau}_2$  is defined as:

$$\tilde{\tau} = \frac{\tilde{\tau}_1}{\tilde{\tau}_2} = \left\langle \frac{\mu_{\tilde{\tau}_1}}{\mu_{\tilde{\tau}_2}}, \frac{\nu_{\tilde{\tau}_2}}{\nu_{\tilde{\tau}_1}} \right\rangle = \langle \mu_M, \nu_M \rangle \quad (11)$$

It is quite possible sometimes, that either  $\mu_{\tilde{\tau}_2}$  or  $\nu_{\tilde{\tau}_1}$ . In such case,  $\tilde{\tau}_1/\tilde{\tau}_2$  cannot be estimated. To handle this situation, the mean value of the attribute for the neighbours is employed.

### 4.3 Algorithm and Illustrative Example

**Input:**  $R$ ; an IF tolerance relation,  $X$ ; dataset with missing attribute values,  $A$ ; set of all attributes  $X$ .

**Output:** The information system with imputed missing data values.

**Begin**

**foreach**  $y \in X$  **and**  $\text{ContainingMissing}(y)$  **do**

$N \leftarrow \text{getNearestNeighbour}(y, k)$

**foreach**  $a \in A$  **and**  $\text{IsMissing}(a(y))$  **do**

$\tau_1 \leftarrow 0, \tau_2 \leftarrow 0$

**foreach**  $z \in N$  **do**

$M \leftarrow ((R \downarrow R_a z)(y) + (R \uparrow R_a z)(y))/2$

$\tau_1 \leftarrow \tau_1 + M * a(z)$

$\tau_2 \leftarrow \tau_2 + M$

**end**

**if**  $(\tau_2 > 0)$

$a(y) \leftarrow (\tau_1/\tau_2)$

**else**

$a(y) \leftarrow \sum a(z)/|N|$

**end**

**end**

**end**

Algorithm 1: Missing Data Imputation using IFRNNI.

The above algorithm work as follows: In a dataset domain, for every instance  $y$ , comprising at least one missing data value for attribute  $a$ , the algorithm obtains its  $k$  nearest neighbours and places them in the set  $N$ . Partial similarities between units are computed by considering the subset of all attributes not missed for the two considered units. For instance, in Example 4.1, the similarity between  $U_0$  and  $U_1$  is determined using attributes  $a_2$  and  $d$ ; between  $U_0$  and  $U_3$ , the attributes  $a_0$ ,  $a_2$ , and  $d$  are used; and between  $U_0$  and  $U_7$ , only the decision attribute  $d$  is used due to missing values in other attributes. This approach ensures that the similarity measure is as comprehensive as possible based on the available data. Thereafter, the missing value are approximated

utilizing  $y$ 's nearest neighbours. Next step is to compute the lower approximation and upper approximation of  $y$  with respect to the instance  $z$ , utilizing the average of these, obtain the final membership  $\mu_M$  and non-membership  $\nu_M$  of the predicted value. The process is conducted for all the instances which belong to  $N$ , and depending upon these calculations over all the neighbours, the algorithm returns a value.

**Example 4.1:** Two datasets are shown in Table 1. The right side of the Table represents the original data with no missing values while the left side represents the same data with some missing attribute values inserted. Missing values are epitomized by “?”. The method of evaluating missing values by IF nearest neighbour algorithm is as follows.

Table 1: Incomplete intuitionistic fuzzy value dataset.

$U$	$a_0$	$a_1$	$a_2$	$d$	$a_0$	$a_1$	$a_2$	$d$
$U_0$	?	?	(0.9, 0.1)	1	(0.6, 0.3)	(0.8, 0.1)	(0.9, 0.1)	1
$U_1$	(0.7, 0.3)	(0.8, 0.2)	(0.7, 0.3)	1	(0.7, 0.3)	(0.8, 0.2)	(0.7, 0.3)	1
$U_2$	(0.7, 0.3)	(0.5, 0.2)	(0.6, 0.3)	2	(0.7, 0.3)	(0.5, 0.2)	(0.6, 0.3)	2
$U_3$	(0.8, 0.2)	?	(0.8, 0.2)	1	(0.8, 0.2)	(0.8, 0.2)	(0.8, 0.2)	1
$U_4$	(0.7, 0.3)	(0.6, 0.2)	(0.6, 0.3)	1	(0.7, 0.3)	(0.6, 0.2)	(0.6, 0.3)	1
$U_5$	(0.6, 0.2)	(0.5, 0.1)	(0.7, 0.3)	2	(0.6, 0.2)	(0.5, 0.1)	(0.7, 0.3)	2
$U_6$	(0.6, 0.3)	?	(0.7, 0.1)	2	(0.6, 0.3)	(0.7, 0.1)	(0.7, 0.1)	2
$U_7$	(0.5, 0.2)	(0.6, 0.2)	?	2	(0.5, 0.2)	(0.6, 0.2)	(0.8, 0.2)	2

In this IF decision system instance  $U_0$  has two missing values  $a_0$  and  $a_1$ , respectively. First, we choose attribute value  $a_0(U_0)$  for imputation.

We calculate Euclidean distance between  $U_0$  and other instances given by Eq. (6). Since attribute value  $a_1(U_0)$  is also missing, so we ignore this attribute at the time of calculating distances and we get the distance between  $U_0$  and  $U_1$  as;

$$d(U_0, U_1) = \sqrt{(0.9 - 0.7)^2 + (0.1 - 0.3)^2 + (0 - 0)^2 + (0)^2} = 0.283$$

Similarly, the distances between  $U_0$  and other instances are calculated.

$$d(U_0, U_2) = 1.068, d(U_0, U_3) = 0.141, d(U_0, U_4) = 0.374,$$

$$d(U_0, U_5) = 1.039, d(U_0, U_6) = 1.039,$$

$$d(U_0, U_7) = \sqrt{\text{equal}(d(U_0), d(U_7))^2} = 1$$

The nearest neighbours of instance  $U_0$  are found in the ascending order of their distances with other instances. Thus,  $N(U_0) = \{U_3, U_1, U_4\}$

To impute  $c_0(U_0)$ , we choose the following three variables  $y, z$  and  $p$  as:

$y = U_0$  is the instance having missing attribute value  $c_0(U_0)$  and  $z, p \in N(y)$ . We first take  $z = U_3$ . On putting third variable  $p = z$  in the formulae of approximation operators, we get no new information. So, we ignore this state and choose value of  $p$  other than  $z$ , either  $U_1$  or  $U_4$ . We calculate the IF tolerance relations by Eq. (1) and all the missing attribute values are ignored in the calculation

First condition,  $y = U_0, z = U_3, p = U_4$ ;  
 $R(y, p) = \min(R_{c_2}(y, p), R_d(y, p))$   
 $= \min(\langle 0.68, 0.16 \rangle, \langle 1, 0 \rangle)$   
 $= \langle \min(0.68, 1), \max(0.16, 0) \rangle = \langle 0.68, 0.16 \rangle$   
 $R_c(p, z) = R_{c_0}(p, z) = \langle 0.92, 0.04 \rangle$

Second condition,  $y = U_0, z = U_3, p = U_1$ ;  
 $R(y, p) = \min(\langle 0.5, 0.26 \rangle, \langle 1, 0 \rangle)$   
 $= \langle \min(0.5, 1), \max(0.26, 0) \rangle = \langle 0.5, 0.26 \rangle$   
 $R_c(p, z) = R_{c_0}(p, z) = \langle 0.92, 0.04 \rangle$

Now, putting the above values in the lower and upper approximation given by Eq.(10), we get

$$R \downarrow R_d z(y) = \inf_{p \in N} I(R(y, p), R_c(p, z))$$

$$= \inf[I(R(U_0, U_1), R_c(U_1, U_3)), I(R(U_0, U_4), R_c(U_4, U_3))]$$

$$= \inf[I(\langle 0.68, 0.16 \rangle, \langle 0.92, 0.04 \rangle), I(\langle 0.5, 0.26 \rangle, \langle 0.92, 0.04 \rangle)]$$

$$= \inf[\langle \min(1, 0.16 + 0.92), \max(0, 0.68 + 0.04 - 1) \rangle, \langle \min(1, 0.26 + 0.92), \max(0, 0.5 + 0.04 - 1) \rangle]$$

$$= \inf[\langle 1, 0 \rangle, \langle 1, 0 \rangle] = \langle \inf(1, 1), \sup(0, 0) \rangle = \langle 1, 0 \rangle$$

$$R \uparrow R_d z(y) = \sup_{p \in N} T(R(y, p), R_c(p, z))$$

$$= \sup[T(R(U_0, U_1), R_c(U_1, U_3)), T(R(U_0, U_4), R_c(U_4, U_3))]$$

$$= \sup[T(\langle 0.68, 0.16 \rangle, \langle 0.92, 0.04 \rangle), T(\langle 0.5, 0.26 \rangle, \langle 0.92, 0.04 \rangle)]$$

$$= \sup[\langle \max(0, 0.68 + 0.92 - 1), \min(1, 1.16 - 0.92, 1.04 - 0.68) \rangle, \langle \max(0, 0.5 + 0.92 - 1), \min(1, 0.26 + 1 - 0.92, 0.04 + 1 - 0.5) \rangle]$$

$$= \sup[\langle 0.6, 0.24 \rangle, \langle 0.42, 0.34 \rangle]$$

$$= \langle \sup(0.6, 0.42), \inf(0.24, 0.34) \rangle = \langle 0.6, 0.24 \rangle$$

Table 2 illustrates the computation of lower and upper approximations using IF T-norm and IF Implication across various attribute pairs.

$$\tilde{r}_1 = \sum_{z \in N} \frac{[R \downarrow (R_d^z(y))] + [R \uparrow (R_c^z(y))]}{2} \times c(z)$$

$$= \frac{(1.0) + (0.6, 0.24)}{2} \langle 0.8, 0.2 \rangle + \frac{(0.96, 0) + (0.84, 0.12)}{2} \langle 0.7, 0.3 \rangle + \frac{(0.96, 0) + (0.84, 0.12)}{2} \langle 0.7, 0.3 \rangle$$

$$= \langle 0.5, 0 \rangle \langle 0.8, 0.2 \rangle + 2 \langle 0.99, 0 \rangle \langle 0.7, 0.3 \rangle$$

$$= \langle 0.4, 0.2 \rangle + \langle 0.906, 0.09 \rangle$$

$$= \langle 0.944, 0.018 \rangle$$

$$\tilde{r}_2 = \sum_{z \in N} \frac{[R \downarrow (R_d^z(y))] + [R \uparrow (R_c^z(y))]}{2}$$

$$= \frac{(1.0) + (0.6, 0.24)}{2} + \frac{(0.96, 0) + (0.84, 0.12)}{2} + \frac{(0.96, 0) + (0.84, 0.12)}{2}$$

$$= \langle 0.5, 0.0 \rangle + \langle 0.99, 0.0 \rangle = \langle 0.99, 0.0 \rangle$$

$$\tilde{r} = \frac{\tilde{r}_1}{\tilde{r}_2} = \left( \frac{\mu_{\tilde{r}_1}}{\mu_{\tilde{r}_2}}, \frac{\nu_{\tilde{r}_1}}{\nu_{\tilde{r}_2}} \right)$$

$$= \left( \frac{0.81}{0.99}, \frac{0}{0.06} \right) = \langle 0.954, 0 \rangle$$

Thus, we get the final predicted value of  $a_0(U_0) = \langle 0.954, 0 \rangle$ .

Table 2:  $a_0(U_0)$  imputation with IFRNNI.

$z$	$p$	$R(y, p)$	$R_c(p, z)$	$I$	$T$	$R \downarrow R_d z(y)$	$R \uparrow R_c z(y)$	$\tilde{r}$
$U_3$	$U_1$	$\langle 0.68, 0.16 \rangle$	$\langle 0.92, 0.04 \rangle$	$\langle 1, 0 \rangle$	$\langle 0.6, 0.24 \rangle$			
$U_3$	$U_4$	$\langle 0.5, 0.26 \rangle$	$\langle 0.92, 0.04 \rangle$	$\langle 1, 0 \rangle$	$\langle 0.42, 0.34 \rangle$	$\langle 1, 0 \rangle$	$\langle 0.6, 0.24 \rangle$	
$U_1$	$U_3$	$\langle 0.92, 0.04 \rangle$	$\langle 0.92, 0.04 \rangle$	$\langle 0.96, 0 \rangle$	$\langle 0.84, 0.12 \rangle$			
$U_1$	$U_4$	$\langle 0.5, 0.26 \rangle$	$\langle 1, 0 \rangle$	$\langle 1, 0 \rangle$	$\langle 0.5, 0.26 \rangle$	$\langle 0.96, 0 \rangle$	$\langle 0.84, 0.12 \rangle$	$\langle 0.954, 0 \rangle$
$U_4$	$U_3$	$\langle 0.92, 0.04 \rangle$	$\langle 0.92, 0.04 \rangle$	$\langle 0.96, 0 \rangle$	$\langle 0.84, 0.12 \rangle$			
$U_4$	$U_1$	$\langle 0.68, 0.16 \rangle$	$\langle 1, 0 \rangle$	$\langle 1, 0 \rangle$	$\langle 0.68, 0.16 \rangle$	$\langle 0.96, 0 \rangle$	$\langle 0.84, 0.12 \rangle$	

## 5 EXPERIMENTAL ANALYSIS

In this section, some experiments are performed on real valued dataset implementing the proposed models and comparison is made with other existing imputation techniques. The impact on the proposed models with variation of parameter k for its different values is investigated. A non- parametric statistical test is also performed for the validation of the results.

### 5.1 Experimental Setup

This subsection describes the datasets used, the other imputation methods used for comparison and also the criteria employed for the comparison. An effective way of estimating imputation methods is that first values are artificially removed from the datasets and then comparison is made between the imputed values produced by the proposed method and the original data values. For this purpose, we have employed 21 datasets from the KEEL dataset repository (Derrac et al., 2015). Table 3 presents the short details of the datasets utilized in the experimentation section. Since none of the datasets include the missing data values, we insert random missing values into them.

Table 3: Description of dataset.

Data Set	#Inst.	#Feat.	#Clas.	Data Set	#Inst.	#Feat.	#Clas.
appendicitis	106	7	2	hepatitis	155	19	2
balance	625	4	3	iris	150	4	3
bands	539	19	2	led-7digits	500	7	10
bupa	345	6	2	monks-2	432	6	2
cleveland	303	13	5	newthyroid	215	5	3
contraceptive	1,473	9	3	spectfheart	267	44	2
ecoli	336	7	8	tae	151	5	3
glass	214	9	7	vowel	990	13	11
haberman	306	3	2	wine	178	13	3
hayes-roth	160	4	3	wisconsin	699	9	2
heart	270	13	2				

Here, MCAR method is used for insertion of missing values in the datasets. For the investigation of the execution of the algorithms under various conditions, we eliminate 5%, 10%, 20% and 30% of the values in the datasets. Perhaps, anything above 30 percent could be too damaging to the data to obtain useful results. A measure is required to compare the results obtained from the imputation algorithms. A commonly used measure to get the difference between the values predicted by a model and the values actually observed in the environment at which experiment is performed, is the Root Mean Square Error (RMSE) (also referred to as root mean square deviation, RMSD). The RMSE of a model being used for prediction with respect to the estimated variable  $Z_{\text{model}}$  is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (z_{obs} - z_{model})^2}{n}}$$

where,  $z_{obs}$  is the observed value and  $z_{model}$  is the imputed value. RMSE measure is used here to compare the yields of the imputation algorithms. Since this measure generates values in different ranges depending upon the ranges of attributes of datasets, we have normalized the employed data with the min-max normalization procedure (Shalabi et al., 2006) so that the comparisons of RMSE values are more practical.

### 5.2 Effect of Parameter K

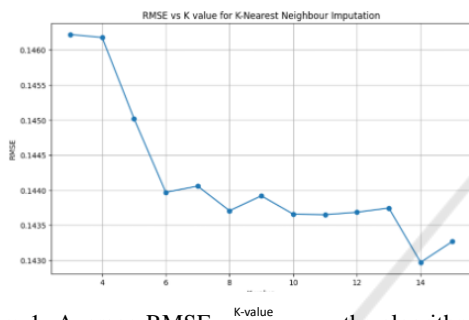


Figure 1: Average RMSE acquired by the algorithm with 5% missing values.

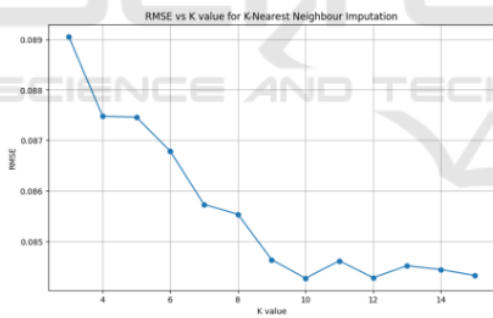


Figure 2: Average RMSE acquired by the algorithm with 10% missing values.

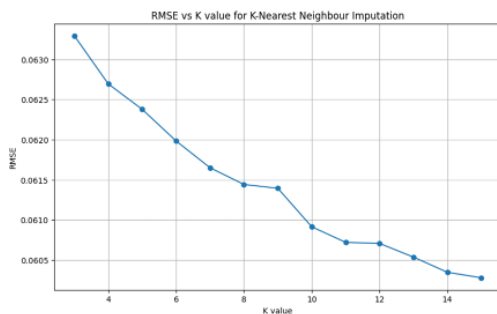


Figure 3: Average RMSE acquired by the algorithm with 20% missing values.

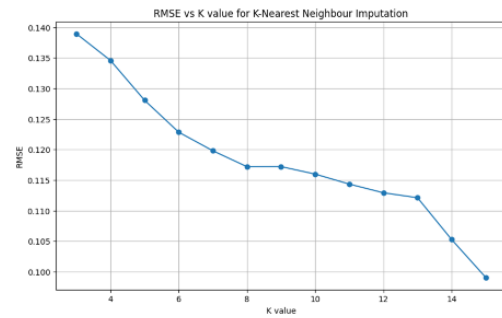


Figure 4: Average RMSE acquired by the algorithm with 30% missing values.

We first begin the experimentation in search algorithms. These parameters are the distance measures, similarity measures, IF t-norms, IF implicators, IF quantifiers, OWA operators and number of neighbours in which most of the best values have been ascertained by other researchers. All these parameters have previously been suggested in section 4. For all other methods the parameters are chosen based on suggestions in (Amiri and Jensen, 2016). The only parameter that needs to be laid down is the number of neighbours, k. To observe the effect of this parameter, we use 21 datasets together with 5%, 10%, 20% and 30% missing values injected into the dataset. The tested values of the parameter k are taken in the range 3 to 15. For the overall convenience, only the average results are given here which are shown in Figures 1-4.

### 5.3 Comparison with Other Missing Data Imputation Methods

In this subsection, a comparison is made between the proposed methods and the other imputation methods. We have compared the proposed methods on 21 datasets with 14 missing value imputation methods using different approaches that are described in introduction section; namely, BPCAI, CMCI, FKMI, KMI, KNNI, LLSI, MCI, SVDI, SVMI and WKNNI, EMI, FRNNI, VQNNI and OWA-FRNNI (Amiri and Jensen, 2016). The average RMSE results of all methods are shown in Figure 5. It can be observed from the figures that for all 5%, 10%, 20% and 30% missing values that the proposed IFRNNI method, have minimum average RMSE values as compared to other imputation methods.



Figure 5: Average RMSE obtained from Missing Data Imputation Algorithms.

Table 4 depicts the obtained results for IFRNNI vs other imputation methods, and it shows that when 5%, 10% and 20% values are missing from the datasets, IFRNNI method has outperformed than most of the imputation methods except FRNNI, VQNNI, OWA-FRNNI, KNNI, BPCAI. The reason is that obtained asymptotic p-values are less than the 0.05 level of significance. For 30% missing data values, IFRNNI has outperformed all other imputation methods.

Table 4: Comparison of the imputation algorithms with IFRNNI in terms of RMSE.

IFRNNI vs	p-value 5%	p-value 10%	p-value 20%	p-value 30%
OWANNI	0.139622	0.121932	0.149178	0.038632
VQNNI	0.121932	0.113770	0.130545	0.035480
FRNNI	0.149178	0.121932	0.149178	0.035480
LLSI	0.073451	0.053725	0.062951	0.024970
KMI	0.022809	0.027306	0.029829	0.015707
SVDI	0.000367	0.000367	0.000321	0.00080
EMI	0.003705	0.000477	0.000702	0.000092
KNNI	0.169775	0.130545	0.149178	0.038362
WKNNI	0.038362	0.035480	0.058186	0.032550
BPCAI	0.149178	0.121932	0.159224	0.079215
MCI	0.004615	0.004615	0.008685	0.004137
CMCI	0.085341	0.073451	0.098741	0.049552
SVMi	0.007838	0.007838	0.024970	0.020812
FKMI	0.032550	0.029829	0.020812	0.005723

## 6 CONCLUSION

Our study introduces novel methods for missing data imputation by integrating IF rough set theory to nearest neighbour approach. This fusion of frameworks offers a comprehensive approach to addressing missing values, leveraging Rough Set Theory’s ability to handle uncertainty and IF set theory’s representation of vagueness. Through empirical evaluations on benchmark datasets, our proposed methods demonstrate superior accuracy and robustness compared to traditional techniques. Notably, our methods offer simplicity and ease of implementation, enhancing their practicality for real-world applications. Overall, this research contributes to advancing missing data imputation methodologies and opens new avenues for leveraging theoretical

foundations to improve data analysis techniques across various domains.

In future work, we will compare the computational efficiency, ease of implementation, interpretability, and generalization capabilities of IFR-NNI with neural network-based imputation methods, focusing on their performance across diverse datasets.

## REFERENCES

Li, Y., Parker, L. E. (2014). Nearest neighbor imputation using spatial-temporal correlations in wireless sensor networks. *Information Fusion*, 15, 64-79.

Sun, Y., Braga-Neto, U., Dougherty, E. R. (2010). Impact of missing value imputation on classification for DNA microarray gene expression data—a model-based study. *EURASIP Journal on Bioinformatics and Systems Biology*, 1-17.

Aydilek, I. B., Arslan, A. (2012). A novel hybrid approach to estimating missing values in databases using k-nearest neighbors and neural networks. *International Journal of Innovative Computing, Information and Control*, 7(8), 4705-4717.

Nelwamondo, F. V., Golding, D., Marwala, T. (2013). A dynamic programming approach to missing data estimation using neural networks. *Information Sciences*, 237, 49-58.

Little, R. J., Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793). *John Wiley & Sons*.

García-Laencina, P. J., Sancho-Gómez, J. L., Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. *Neural Computing and Applications*, 19, 263-282.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ..., Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6), 520-525.

Batista, G. E., Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence*, 17(5-6), 519-533.

Grzymala-Busse, J. W., Grzymala-Busse, W. J. (2005). Handling Missing Attribute Values. *Data Mining and Knowledge Discovery Handbook*, 37.

Kim, H., Golub, G. H., Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2), 187-198.

Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of climate*, 14(5), 853-871.

Oba, S., Sato, M. A., Takemasa, I., Monden, M., Matsubara, K. I., Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16), 2088-2096.



- Honghai, F., Guoshun, C., Cheng, Y., Bingru, Y., Yumei, C. (2005). A SVM regression-based approach to filling in missing values. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (pp. 581-587). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Li, D., Deogun, J., Spaulding, W., Shuart, B. (2004). Towards missing data imputation: a study of fuzzy k-means clustering method. In *Rough Sets and Current Trends in Computing: 4th International Conference, RSCTC 2004, Uppsala, Sweden, June 1-5, 2004. Proceedings 4* (pp. 573-579). Springer Berlin Heidelberg.
- Liao, Z., Lu, X., Yang, T., Wang, H. (2009). Missing data imputation: a fuzzy K-means clustering algorithm over sliding window. In *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery* (Vol. 3, pp. 133-137). IEEE.
- Sharpe, P. K., Solly, R. J. (1995). Dealing with missing values in neural network-based diagnostic systems. *Neural Computing & Applications*, 3, 73-77.
- Bengio, Y., Gingras, F. (1995). Recurrent neural networks for missing or asynchronous data. *Advances in neural information processing systems*, 8.
- Pyle, D. (1999). Data preparation for data mining. *Morgan Kaufmann*.
- Amiri, M., Jensen, R. (2016). Missing data imputation using fuzzy-rough methods. *Neurocomputing*, 205, 152-164.
- Pereira, R. C., Santos, M. S., Rodrigues, P. P., Abreu, P. H. (2020). Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes. *Journal of Artificial Intelligence Research*, 69, 1255-1285.
- Huang, S. Y. (Ed.). (2013). *Intelligent decision support: handbook of applications and advances of the rough sets theory*.
- Pawlak, Z. (2012). *Rough sets: Theoretical aspects of reasoning about data* (Vol. 9). Springer Science & Business Media.
- Atanassov, K. T. (1999). Intuitionistic fuzzy sets (pp. 1-137). Physica-Verlag HD.
- Bustince, H., Burillo, P. (1996). Vague sets are intuitionistic fuzzy sets. *Fuzzy sets and systems*, 79(3), 403-405.
- Cornelis, C., De Cock, M., Kerre, E. E. (2003). Intuitionistic fuzzy rough sets: at the crossroads of imperfect knowledge. *Expert systems*, 20(5), 260-270.
- Jensen, R., & Cornelis, C. (2011). Fuzzy-rough nearest neighbour classification. In *Transactions on rough sets XIII* (pp. 56-72). Springer Berlin Heidelberg.
- Szmidt, E., & Kacprzyk, J. (2001). Entropy for intuitionistic fuzzy sets. *Fuzzy sets and systems*, 118(3), 467-477.
- Derrac, J., Garcia, S., Sanchez, L., Herrera, F. (2015). Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *J. Mult. Valued Logic Soft Comput*, 17, 255-287.
- Al Shalabi, L., Shaaban, Z., Kasasbeh, B. (2006). Data mining: A preprocessing engine. *Journal of Computer Science*, 2(9), 735-739.