Fuzzy C-mean Missing Data Imputation for Analogy-based Effort Estimation

Ayman Jalal AlMutlaq[1], Dayang N. A. Jawawi[2], Adila Firdaus Binti Arbain[3]
Department of Software Engineering, School of Computing, Faculty of Engineering
,Universiti Teknologi Malaysia, Johor Bahru, Malaysia[1, 2, 3]

***Abstract*—The accuracy of effort estimation in one of the major factors in the success or failure of software projects. Analogy-Based Estimation (ABE) is a widely accepted estimation model since its flow human nature in selecting analogies similar in nature to the target project. The accuracy of prediction in ABE model in strongly associated with the quality of the dataset since it depends on previous completed projects for estimation. Missing Data (MD) is one of major challenges in software engineering datasets. Several missing data imputation techniques have been investigated by researchers in ABE model. Identification of the most similar donor values from the completed software projects dataset for imputation is a challenging issue in existing missing data techniques adopted for ABE model. In this study, Fuzzy C-Mean Imputation (FCMI), Mean Imputation (MI) and K-Nearest Neighbor Imputation (KNNI) are investigated to impute missing values in Desharnais dataset under different missing data percentages (Desh-Miss1, Desh-Miss2) for ABE model. FCMI-ABE technique is proposed in this study. Evaluation comparison among MI, KNNI, and (ABE-FCMI) is conducted for ABE model to identify the suitable MD imputation method. The results suggest that the use of (ABE-FCMI), rather than MI and KNNI, imputes more reliable values to incomplete software projects in the missing datasets. It was also found that the proposed imputation method significantly improves software development effort prediction of ABE model.**

Keywords*—Analogy-based effort estimation; imputation; missing data; fuzzy c-mean

I. Introduction

Software development effort is considered one of the most significant metrics estimated in software projects due to the reasons that planning, developing, managing and all other important aspects of project depend extremely on accurate estimation of development effort[1]. Many effort estimation models have been introduced by researchers in software engineering domain , they can be classified into two major categories: first is parametric models which depend on statistical analysis of software projects data and assumed a linear relationship between effort and other project attributes, and second is Machine Learning (ML) models which depends on soft computing and artificial intelligence methods and assumed a non-linear relationship between effort and other project attributes [2, 3]. Among many ML models Analogy-Based Estimation (ABE) is a widely accepted estimation model since its flow human nature in selecting analogies similar in nature to the target project[4].

Missing data (MD) in software engineering datasets is major problem that affects the performance of effort prediction models [5, 6]. Many techniques are proposed to solve this problem includes : deletion, toleration, and imputation of missing data [7]. Missing data imputation is the most investigated technique in software effort estimation and KNN imputation was the popular adopted method [8].

Almutlaq and Jawawi [9], classified missing data imputation challenges for software effort estimation into two major categories, the categories are performance oriented and dataset challenges. Performance oriented challenges refers to challenges and issues that exist within the techniques itself on a performance level (missing data Accuracy, Model performance accuracy, and time efficiency). While the dataset challenges revolve around the role of the dataset and its effect on the missing data imputation techniques (numerical data imputation, categorical data imputation, dataset characteristics and size variety, and MD Mechanism Variety).

MI and KNNI are the most prominent missing data imputation techniques that have been used for ABE model [8]. MI method is considered as static imputation without analyzed the dynamic nature for each missing case in the feature concerned [10, 11]. KNNI depends on neighbor cases which may be related or not to the missing project values and derived a dynamic imputation value for each missing case for the feature concerned in the uncompleted dataset[12].

Identification of the most similar donor values from the completed software projects dataset for imputation is a challenging issue in the existing missing data techniques adopted for ABE model. Clustered completed software projects into homogeneous clusters based on the selected dataset attributes, and then identify more reliable donors cases to the incomplete project to impute missing values based on clustered data have not been yet investigated in ABE domain.

This study concerns on improve the performance of ABE model through adopting a new imputation method based on FCM technique. And compare empirically the results with KNN imputation and Mean Imputation (MI) for ABE model using different missing ratio of MNAR missingness mechanism.

Rest of the paper is organized as follow. Section II presents the concepts of ABE model, missing data, and techniques for handling missing data in software engineering datasets. Section III presents the concept of Fuzzy C-Mean clustering. Section IV presents related research studies for missing data techniques in software engineering domain and ABE model. Section V presents the proposed (ABE-FCMI) imputation technique. Section VI presents empirical evaluation design employed in this study. Section VII presents and discusses the

reported results. Section VIII discusses internal and external threats to validity for this research study. Section IX concludes research findings and gives direction for some future work.

II. Background

This section presents the concepts of analogy-based effort estimation, missing data, and fuzzy c-mean (FCM) clustering.

A. Anolgy-Based Estimation (ABE)

Analogy based estimation proposed by Shepherd and Schofield as one of the most prominent non-algorithmic effort estimation model [13] .Comparison dependent process of comparing similar projects to the target project is done in order to derive the development effort in ASEE. Similarity measures are used to determine similar projects. Simplicity and estimation capability make it a widely accepted model in software effort estimation field. ABE consist of four parts:

- Historical completed software engineering projects dataset.
- Determine the level of similarity through Similarity Function.
- Estimate the software development effort by considering the similar projects found by the similarity function through solution function.
- Associated retrieval rules

The estimation process of ABE is accomplished in the following stages:

- A historical dataset in constructed based on the collected information of previous projects.
- For a comparison purpose select attributes are chosen.
- Retrieve similar projects to the target project based on the selected similarity function.
- Estimate the target project effort based on the selected solution function.

Similarity Function: Level of similarity between two projects is determined through similarity function that compares the attributes of both projects. Euclidian Similarity (ES) and Manhattan Similarity (MS) are two common similarity functions. (ES) function is represented in Equation 1.

$$Sim(p,p') = \frac{1}{\left|\sqrt{\sum_{i=1}^{n} w_i Dis(fi,fi') + \delta}\right|} \quad \delta = 0.0001$$

$$Dis(f_1,f_2) = \begin{Bmatrix} (f_1 - f_2') \; if \; f_1 \; and \; f_2' \; are \; numerical \; or \; ordinal \\ 0 \; if \; f_1 \; and \; f_2' are \; nominal \; and \; f_1 = f_2' \\ 1 \; if \; f_1 \; and \; f_2' are \; nominal \; and \; f_1 \neq f_2' \end{Bmatrix} \quad (1)$$

Where projects in comparison are p and p' whereas Wight given to each attribute as wi. wight range between 0 and 1. The ith attribute of each project represented as fi and fi' and n represent the number of attributes. For gain none zero result δ is used. Solution Function: To derive software effort estimation based on most similar projects defined by similarity function a solution function is applied. Most dominant used solution functions are: inverse distance weighted mean [14] , closest analogy as the most similar project [15] , average of most similar projects [13] , median of most similar projects [16]. The median value of effort gained from K most similar projects, as K>2, described by Median. The average value of efforts gained from K most similar projects, as K>1, is described by Average.

B. Missing Data Concept

Missing data (MD) problem is a major challenge in software engineering datasets. Accurate software effort estimation depends strongly on the quality of datasets used for estimation process. In this subsection MD mechanisms and MD techniques (treatments) are elaborated.

C. Mechanisms of Missing Data

Missing data mechanisms are assumptions about the type and distribution of missing values [17].This identification of missing mechanism identify the missing treatment to be applied [7]. Three type of missing data mechanism are identified.

First Missing Completely At Random (MCAR) MD are independent of any variable observed in the data set, second Missing At Random (MAR) means that the MD may depend on variables observed in the data set, but not on the MD themselves, third (MNAR) in which the MD depend on the MD themselves and not on any other observed variable.

D. Techniques for Missing Data

Missing data treatment can be grouped in three methods as first MD deletion, second MD toleration, and third MD imputation.

MD ignoring (deletion) in this technique it simply handle the missing values by deleting them. MD deletion is properly suitable when the percentage of missing data is low. It is not utilize when consecutive data is missing like NIM (MNAR) mechanism [7, 18]. MD toleration in this method the missing value is assigned a NULL value and did not deleted from the dataset and the analysis is performed to same data [18]. MD imputation MD imputation method is employed to fill up the missing values and reaches a complete data set so that later this dataset can be utilized in enhancing the estimation of software development effort. KNN imputation is the most prominent method of imputation in software effort estimation [8, 19, 20].KNN provides a good result so far because it dost follow explicit mechanisms. Euclidean Distance and Manhattan Distance is used as a similarity measure to find nearest neighbors in KNN imputation methods.

III. Fuzzy C-Mean (FCM) Clustering

KNNI uses whole completed dataset for identifying similar neighborhood donor cases based on some distance measure, for ABE context it is important that donor cases to incomplete projects are come from similar projects in characteristics and nature to incomplete software project to impute missing values.

Clustering strategy as a data mining technique has been utilized recently to impute missing value. The idea behind using clustering in MD imputation is to impute incomplete record missing values from similar cluster that incomplete

record located in, accuracy of imputation is improved by clustering data to groups with the same similarity features so that the range to substitute missing values is within cluster scope[21].

Clustering techniques can be divided into two major categories, hard clustering and soft (fuzzy) clustering. In hard clustering techniques, data object is belong to only one cluster which is the most similar cluster , however in fuzzy clustering a dataset object is belong to each one of clusters with a certain similarity given by membership function [22].

Hard clustering imputation techniques has been employed by many researchers such as k-means [23-25] in which incomplete data object missing values is imputed based on cluster information it is belong to. However in case of missing dataset there is uncertainty of incomplete data object is belonging definitely to certain cluster, so the need for fuzzy clustering imputation methods have been introduced such as FCMI [26-28] . The intra-variance in clusters is decreases by FCM compared to k-means algorithm [29] , moreover FCM is less sensitive to stuck on local minimum situation because of continuous membership function values [30]. Fuzzy imputation achieved higher performance compared to hard clustering imputation as denoted in experimental results [31].

Zadeh introduced the concept of fuzzy logic [23, 32]. Fuzzy logic is a computation approach based on degree of truth to represent uncertainty concept in information. Fuzzy theory and fuzzy set are introduced to solve the problem of imprecise information and uncertainty in missing data. Fuzzy capabilities are utilized to find plausible imputation values [31, 33, 34].

One dataset element can belong to two or more subsets in fuzzy clustering rather than crisp clustering. In FCM one dataset element can belong all clusters with different membership value associated to each clusters [35, 36].

Fuzzy C-Means (FCM) adopted recently in solving missing data problem [27, 28, 37]. Missing value can be derived by the calculated distance from clustered complete dataset based on obtained membership values.

This study focus on missing data imputation by clustering the completed projects into several clusters where they have similar connection between the features subsets.to best of our knowledge no research study has adopted FCM for ABE model.

FCM is a form of iterative algorithm. The goal of FCM is to find cluster centers (centroids) that minimize objective function (dissimilarity).The dissimilarity function (J) which is used in FCM is given Equation 2.

$$J = \sum_{i=1}^{n}\sum_{j=1}^{c} \mu_{ij}{}^{m} d_{ij}{}^{2} \quad (2)$$

μij is a membership function for i-th observation of the jth centroid, where $\sum_{i=1}^{n} \mu_{ij} = 1$

c is the number of clusters.

n is the number of observations.

dij is the Euclidian distance (||Xi-Cj||2) between ith centroid(ci) and jth observation.

m is the fuzzy degree ,m=2 is the general used value.

The cluster center (centroid) rj of jth cluster is given using equation 3.

$$rj = \frac{\sum_{j=1}^{c} \mu_{ij}{}^{m} xi}{\sum_{j=1}^{c} \mu_{ij}{}^{m}} \quad (3)$$

Compute the Euclidian distance and Update membership function μi j using equation 4.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \quad (4)$$

The FCM algorithm can be elaborated as follow:

Algorithm 1: FCM Algorithm

REQUIRE: Input data to be clustered (X1, X2, , Xn). 2. Number of clusters (c), fuzzy degree value (m), maximum number of iterations allowed (I), the smallest desired error(ε),initial objective function (J0 = 0).
Step 1: Begin
Step 2: Initialize randomly membership function to each observation (μi j)
Step 3: Calculate centroid (cluster center) (rj) using equation 3
Step 4: Calculate the Euclidean distance, update the membership function (μi j) using equation 4
Step 5: Calculate objective function using equation 2
Step 6: Check for convergence criterion
IF (‖Ji − J(i − 1)‖) < ε OR (i > I) , then stop the process.
ELSE repeat step 2 to 6 until maximum iteration reached.
Step 7: END

IV. Related Work

The quality of past software dataset projects play major role in the performance of ABE model since it depend on historical past projects to predict the effort of target project. Researchers investigated missing data treatment techniques wildly in software engineering filed but few concentrate on ABE model. Idri, et al. [8] conducted a systematic mapping study in software engineering domain reviewed existing techniques treating missing data, it have been found that missing data imputation is the most used approach and KNN imputation is the most adopted method. Huang, et al.[6] Evaluated empirically data preprocessing techniques used for machine learning effort estimation models; the study validated missing data treatment techniques effectiveness to improve accuracy of prediction effort. Almutlaq and Jawawi [9] Reviewed recent missing data techniques in software effort estimation field, the study elaborated two major challenges that are imputation technique performance oriented and incomplete dataset oriented.

Strike, et al.[5] Investigated three missing data techniques (deletion, mean imputation, and hot-deck imputation) with three missing mechanisms (MCAR, MAR,and NIM) on regression effort estimation model.it have been found that hot-deck imputation outperformed other methods. Cartwright, et al.[19] Founded that KNN imputation has better results than mean imputation and missing data toleration in regression effort estimation model for MCAR missing data mechanism. Twala and Cartwright [20] combined KNN imputation with multiple imputation approach for Decision Trees effort estimation model, experimental results improved predictive accuracy of effort estimation using the proposed ensemble method. Sentas and Angelis [38] Investigated multinomial logistic regression (MLR) imputation for categorical missing data type in ISBSG dataset, the accuracy of regression estimation model improved especially with the case of high percentage of missing values. Li, et al.[18] Studied the relation between percentage of missing data (MCAR missing mechanism) and accuracy of AQUA model (form of ABE), the results confirmed a positive quadratic relation between percentage of missing data and accuracy of effort prediction. Song, et al.[7] Analyzed the impact of missing percentage and messing mechanisms on the accuracy of C4.5 effort estimation model using toleration and KNN imputation methods, the accuracy of prediction is severely affected in cases missing percentage above 40%. Idri, Abnane et al. [39] Conducted a study to evaluate prediction accuracy of ABE using different missing data techniques (toleration ,deletion ,and KNN imputation) with all missing mechanisms ,KNN imputation had superior improvement in ABE performance results.

Abnane and Idri [40] Investigated MD techniques (toleration, deletion, and KNN imputation) under different missing ratios and MD mechanisms for Fuzzy-ABE model using PRED (0.25) and SA as accuracy measures, they found that SA and PRED(0.25) measured different characteristics of technique performance. Huang, Li et al [41] Investigated data-preprocessing techniques (MD, normalization, feature selection) for ABE model under ISBSG dataset, KNNI improved ABE performance significantly compared to MI. Idri, Abnane et al [42] proposed SVR (Support Vector Regression) imputation, empirical results indicated that SVRI outperformed KNNI under different missing ratio and MD mechanisms for ABE model. Abnane and Idri [43] investigated mixed (Numerical and categorical) MD imputation techniques for ABE model, imputation techniques achieved better accuracy results, there is no significant difference between SVR and KNNI for mixed MD imputation. Muhammad Arif Shah [44] proposed Median Imputation of the Nearest Neighbor (MINN) for ABE mode , the investigation of the proposed model under Desharnais dataset outperformed both MI and KNN under MNAR mechanism.

Abnane, Hosni et al. [45] optimize parameters of KNN imputation using grid search, the optimize KNN imputation improved ABE significantly compared with regular KNN imputation. Abnane, Idri et al. [46] Proposed 2FA-KP-I (Fuzzy Analogy k-Prototypes Imputation) to impute mixed MD in ABE model, 2FA-KP-I outperformed KNNI under different missing ratio and MD mechanisms for ABE in the studied datasets.

Table I introduced literature review of MD techniques used in ABE model, it also summarized the type of MD, imputation methods used MD mechanism, and the findings for each study. As can be seen from Table I that KNNI and MI is the most used techniques. Literature review in Table I gives indication that the increased MD ratio negatively affected ABE performance, and MNAR MD mechanisms significantly decreased ABE performance.

MI method impute fixed value for all missing data in the same column (feature),this is done by replacing all missing value with the average value of the feature concerned. MI method is considered as static imputation without analyzed the dynamic nature for each missing case in the feature concerned, MI can alter the variance of the data and the relationships between variables does not preserved like correlation [10, 47, 48].

KNNI depends on neighbor cases of the missing value and derived a dynamic imputation value for each missing case for the feature concerned. KNN imputation have limitations related to : first not efficient for large dataset size ,second it imputes values based on the neighbors which may or may not be the related projects for donor values, third depend on parameter setting for KNN algorithm , and fourth KNNI performance is decreased with MNAR missingness mechanism [12, 39, 49, 50].

As can be seen from literature identification of the most similar donor values from the completed software projects dataset for imputation is a challenging issue in the existing missing data techniques adopted for ABE model. Clustered completed software projects into homogeneous clusters based on the selected dataset attributes, and then identify more reliable donors cases to the incomplete project to impute missing values based on clustered data have not been yet investigated by most researchers in ABE domain.

TABLE I. LITERATURE REVIEW OF MD TECHNIQUES IN ABE MODEL

Reference	Type of MD	Imputation Method	MD Mechanism
[18]	Numerical, Categorical	Toleration	MCAR
Finding	The results indicate that increased percentage of MD affected negatively accuracy prediction of AQUA (type of ABE model). The study suggested 40% upper limit of MD to get acceptable accuracy results of AQUA. The study suggested increased historical projects and attributes in the studied datasets to get better accuracy results of AQUA as MD percentage increased.		
[39]	Numerical	Toleration, Deletion and KNN imputation	MAR MCAR MNAR
Finding	KNN imputation improved ABE accuracy results compared to toleration or deletion of MD. The results shown that as the percentage of MD increased the accuracy of ABE is decreased .The results founded that the missingness mechanism affect the performance of ABE, accuracy of ABE is decreased significantly under MNAR compared to both MAR and MCAR.		
[40]	Numerical	Toleration, Deletion and KNN imputation	MAR MCAR MNAR
Finding	Fuzzy-ABE model have been got more accurate results using KNNI compared to deletion or toleration. PRED (.25) accuracy result confirmed SA measure. The results suggested to combine SA with other accuracy measure.		
[41]	Numerical	Mean imputation (MI) ,KNN imputation	Original missing values in ISBSG dataset
Finding	The investigated experimental results on ISBSG dataset concluded that KNN imputation as significant part of data-preprocessing stage improved the accuracy results of ABE compared to MI.		
[42]	Numerical	Support vector regression (SVR) imputation, KNN imputation	MAR MCAR MNAR
Finding	SVR imputation outperforms KNN imputation for both classical and fuzzy analogy effort estimation. The results shown that SVR imputation is less sensitive regarding MD percentage compared to KNN imputation. The results confirmed that for both SVR imputation and KNN imputation had worse performance under MNAR mechanism compared to both MAR and MCAR.		
[43]	Numerical and categorical MD	toleration, deletion, KNNI, SVR imputation	MAR MCAR MNAR
Finding	The results confirmed that imputation techniques achieved better accuracy improvements compared to toleration and deletion. In term of SA accuracy measure there is no significant difference between SVR and KNNI for mixed MD imputation. MNAR mechanism significantly affects ABE accuracy results for mixed MD imputation.		
[44]	Numerical	KNNI , MI , Median Imputation of the Nearest Neighbor (MINN)	MNAR
Finding	Experimental results reported that MINN outperformed both KNNI and MI for the studied Desharnais dataset. The results confirmed that there is no significant difference in accuracy improvement between KNNI and MINN due to the small size of the studied dataset. To generalize accuracy results there is a need to investigate large size datasets.		
[45]	Numerical	GS(Grid Search)-KNNI , E(Ensemble)-KNNI ,UC(Uniform Configuration)-KNNI	MAR MCAR MNAR
Finding	The proposed E-KNNI employed parameter optimization at imputation step. The results indicate that E-KNNI accuracy outperform GS-KNNI. E_KNNI and GS-KNNI had similar accuracy results. For MNAR mechanism E-KNNI significantly outperforms GS-KNNI.		
[46]	Numerical and categorical MD	2FA-KP-I (Fuzzy Analogy k-Prototypes Imputation), KNNI	MAR MCAR MNAR
Finding	The results found that 2FA-KP-I outperforms KNNI on four software engineering datasets under different missing ratio and MD mechanisms. Mean standard error (RMSE) is considered as imputation accuracy measure to evaluate competitive imputation techniques. The results indicate that MD mechanisms affected imputation accuracy for both 2FA-KP-I and KNNI, MNAR mechanism had significant impact on both.		

V. PROPOSED (ABE-FCMI) IMPUTATION TECHNIQUE

This section discusses the proposed (ABE-FCMI) imputation technique for imputing software engineering datasets. (ABE-FCMI) employed fuzzy clustering to divide the completed software projects into homogeneous clusters based on their features. Group completed data into similar features using FCM is the main operation to get for each feature the centroid value and obtain cluster centers finally.

The proposed (ABE-FCMI) method tries to solve gaps of, first selecting proper adjacent cases to derive the final missing data estimation value, and second improve ABE performance through MD imputation of MNAR missingness mechanism.

The basic idea behind using (ABE-FCMI) technique in ABE context is to impute incomplete software projects missing values based on homogeneous clustered completed software projects with high similarity within cluster and dissimilar with software projects in other clusters. Identification of similar donor cases for imputation is then assessed based on incomplete project membership values on each cluster.

In this study the idea of FCMI is borrowed from literature [27, 33] and applied to the problem of MD in ABE model to improve the prediction accuracy of software effort estimation.

The algorithm of the proposed (ABE-FCMI) method is as follow:

Algorithm 2: ABE - FCMI Algorithm

REQUIRE: Normalize the software projects dataset (D) using min-max normalization. Separate dataset (D) into two subsets: Complete software projects dataset (DC) and Incomplete software projects dataset (DM).
Step 1: **Begin**

Step 2: For all Complete software projects dataset (DC):
i. Calculate the cluster center (centroid) using Equation 3.
ii. Compute the Euclidean distance
iii. Update the membership function using Equation 1, 2, and 3.

Step 3: For all Incomplete software projects dataset(DM):
i. Calculate membership function to cluster centers that
are Calculated from step 2.

Step 4:For each incomplete software project calculate imputation value using membership value calculated from step 3 and cluster centers calculated from step 2.

Step 5 : End

The proposed (ABE-FCMI) algorithm imputes each incomplete project using information about membership function and the calculated cluster centers of completed projects. Generating of missing values using particular missingness mechanism and normalization of the dataset is taken in advanced before the imputation process started.

The processes of the proposed (ABE-FCMI) imputation method for ABE model is shown in Fig. 1 which include mainly : calculate cluster centers of complete software projects, calculate membership values for each incomplete software project, and estimate the imputed missing values. In first step the whole dataset is separated to complete and incomplete datasets. Cluster centers for complete software projects are calculated using FCM algorithm. In second step for each incomplete software project the membership values to given cluster center are calculated. In third step the imputation value is estimated based on membership values of incomplete software project calculated in second step and the cluster centers of complete software projects calculated in first step. The imputed dataset is used to evaluate the accuracy of prediction of ABE model as elaborated in Fig. 1.

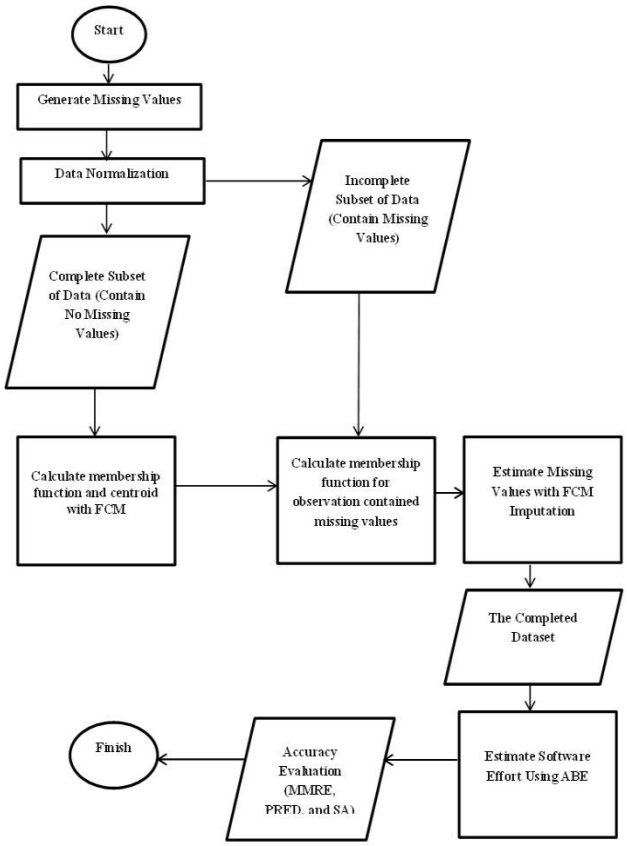

Fig. 1. The Proposed ABE- FCMI Method.

VI. Empirical Evaluation Design

In this section the empirical evaluation design is elaborated to define: first the datasets used in this study, second performance accuracy measures used to assess ABE prediction results, and third the adopted empirical process employed in this study.

A. Data Sets Description

Desharnais dataset as one of the most common datasets in the field of software effort estimation [51]. Recent research studies investigate Desharnais dataset imputation for ABE performance evaluation [39, 42, 44]. The data contain 81 software projects related to Canadian Software Company, 77 projects are complete with no missing values, and four projects are considered incomplete with some missing values. The data has nine features, all features are numerical except one feature which are language that are categorical. Effort feature is considered as dependent feature and other features are considered as independent features. The statistical details of Desharnais dataset is given in Table I. In projects number 38, 44, the TeamExp feature values are missing. In projects number 38, 66, and 75, the ManagerExp feature values are missing. The Histogram and pattern of missing data for Desharnais dataset can be seen in Fig. 2.

TABLE II. DESHARNAIS DATASET DESCRIPTION

Feature	Description	Min	Max	Mean	Std Dev
Effort	Development Effort in person-hours	546	23940	4923.516	4646.751
TeamExp	Team Experience in Years	0	4	2.244	1.331
ManagerExp	Manager Experience in Years	0	7	2.803	1.47
Length	Length of Project in months	1	39	11.716	7.4
Transections	Number of Transactions	9	886	179.901	143.315
Entities	Number of Entities	7	387	122.726	86.178
PointsAdjust	Number of Adjusted Function Points	73	1127	311.014	189.185
Envergure	Function Point Complexity Adjustment factor	5	52	27.014	10.851
PointsNonAdjust	Project Size Measured In Unadjusted Function Points. (Entities Plus Transactions)	62	1116	295.765	197.937

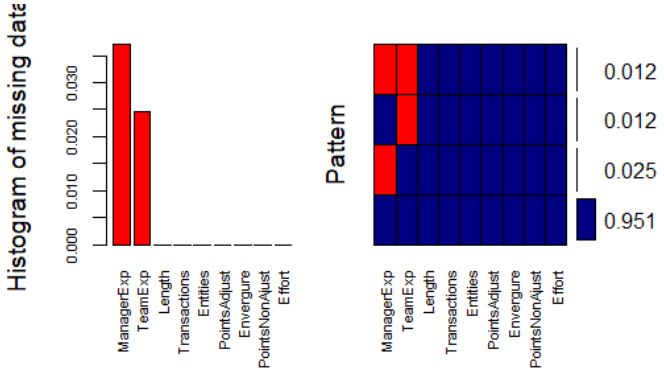

Fig. 2. Missing Data Histogram and Patterns for Desharnais Dataset.

The percentage of missing values in Desharnais dataset is relatively very low. In this study tow Desharnais datasets with different missing ratio are artificially created with MNAR missing mechanism to validate proposed missing data imputation methods for ABE model. Desh-Miss1 dataset 28.395% missing row ratio (23 out of 81 projects have missing values) and 3.33 % missing cell ratio (24 missing cells out of 720 cells) with MNAR missingness mechanism, and Desh-Miss2 dataset with 69.135 % missing row ratio (56 out of 81 projects have missing values) 7.916 % missing cell ratio (57 missing cells out of 720 cells) with MNAR missingness mechanism. Artificial missing data generation in software effort estimation has been performed in studies such as [18, 39]. The Histogram and pattern of missing data for Desh-Miss1 and Desh-Miss2 datasets can be seen in Fig. 3 and Fig. 4, respectively.

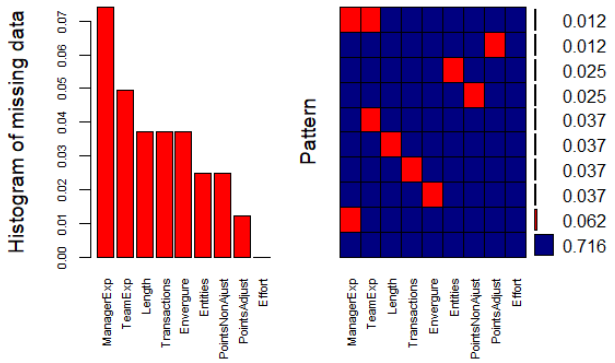

Fig. 3. Missing Data Histogram and Patterns for Desh-Miss1 Dataset.

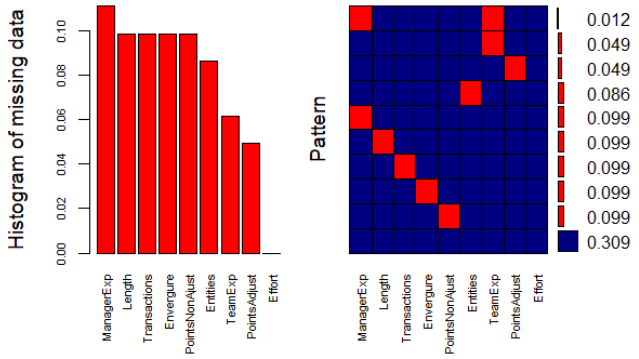

Fig. 4. Missing Data Histogram and Patterns for Desh-Miss2 Dataset.

B. Performance Accuracy Metrics

Several metrics have been used to evaluate the performance of estimation models which include Mean Magnitude of Relative Error (MMRE) measure that based on Relative Error (RE), and Magnitude of Relative Error (MRE) [13]. MMRE as most used evaluation metrics is defined as:

$$RE = (Estimated - Acutal)/Actual) \quad (6)$$

$$MRE = |Estimated - Actual|/(Actual) \quad (7)$$

$$MMRE = \sum_{i=1}^{N} MRE/N \quad (8)$$

Percentage of the prediction (PRED) is defined as:

$$PRED(X) = \frac{A}{N} \quad (9)$$

Where, A is the number of projects with MRE less than or equal to X and N is the total number of test set projects. Most effort estimation models are compared within X is 0.25 as acceptable value [52]. Shepperd and MacDonell [53] proposed SA measure that based on mean absolute error (MAE). SA considered as unbiased and standardized accuracy measure and gives an idea about the effectiveness of estimation model compared to random guessing.

$$MAR = \frac{\sum_{i=1}^{N} AEi}{N} \quad (10)$$

$$SA = 1 - \frac{MARp_i}{MARp_0} \quad (11)$$

Where MARp_i is the Mean Absolute Error of estimation technique p_i , and MARp_0 is the mean of a large number of

random guesses (in our case 1000). The goal of estimation model is to minimize MMRE and maximizes PRED and SA prediction results for software effort estimation models.

Cross validation: Cross-Validation is introduced to give a more realistic accuracy evaluation to the estimation model. By dividing the historical dataset into multiple training and testing sets. These groups have almost equal size, one group is selected as test group and the remaining groups will be test groups. After that the estimation is computed for the test set and iteratively the process will be continued until all set are involved in the estimation , this depend of the number of sets. This insures the verification of all projects. Actually, all the projects are considered as a test case only once in all iterations. The final performance achieved from all the iterations is considered as mean value of performance metrics. MMREs, PREDs, and SAs mean values from all iteration is considered as MMRE, PRED, and SA final value.

C. Empirical Process

The empirical process adopted for this study is presented in fig. 5. As can be seen from Fig. 5, it is consists of four main steps: generating missing values, missing data imputation, ABE effort estimation, and accuracy evaluation. The design for the used empirical process followed similar approach used in [18, 39, 44] for evaluating the impact of MD imputation for ABE performance prediction.

Step 1: Generate missing values: in this study tow Desharnais datasets with different missing ratio are artificially created with MNAR missing mechanism to validate proposed missing data imputation methods for ABE model. Desh-Miss1 dataset with 28.395% missing row ratio (23 out of 81 projects have missing values) and 3.33 % missing cell ratio (24 missing cells out of 720 cells) with MNAR missingness mechanism, and Desh-Miss2 dataset with 69.135 % missing row ratio (56 out of 81 projects have missing values) 7.916 % missing cell ratio (57 missing cells out of 720 cells) with MNAR missingness mechanism. Artificial missing data generation in software effort estimation has been performed in studies such as [18, 39]. The Histogram and pattern of missing data for Desh-Miss1 and Desh-Miss2 datasets can be seen in Fig. 3 and Fig. 4 respectively. Table IV of Appendix presents a sample of the outcome (Desh-Miss2) of this step using MNAR mechanism with 69.135 % of MD on Desharnais dataset. Step 2: Missing data imputation: three imputation techniques (MI, KNNI, and (ABE-FCMI)) are used to impute missing values. The performances of these techniques are compared later to identify best imputation technique adopted for ABE prediction. Table XV of Appendix presents the outcome of the Step 2 using (ABE-FCMI) imputation under MNAR mechanism at 69.135% of MD on the sample data of Table XV. Step 3: Effort Estimation using ABE: software development effort using ABE model is predicted from the imputed dataset (complete dataset).Euclidian distance is used as similarity function and mean is used as solution function in ABE algorithmic procedure. Step 4: Accuracy evaluation: The performance of ABE is evaluated after each imputation technique to discover which imputation method outperforms the other. MMRE, PRED (0.22), and SA are used as accuracy estimation measures. Three-fold cross-validation is considered as evaluation method in ABE prediction model.

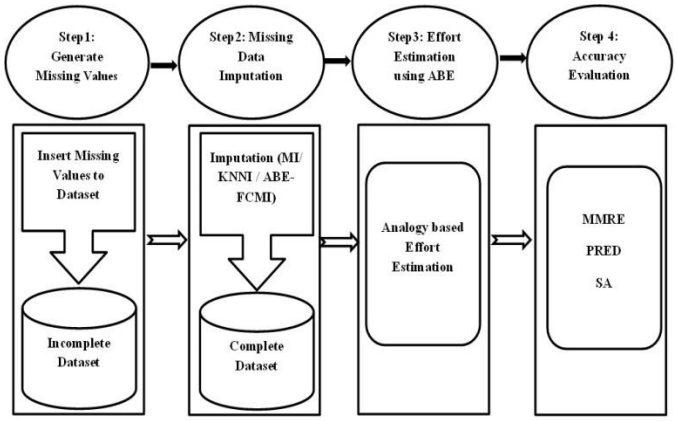

Fig. 5. Empirical Process for (MI, KNNI, and ABE-FCMI) Imputation Methods for ABE Prediction Model.

VII. Result and Discussion

This section presents the experimental results for evaluating ABE performance using three imputation methods (MI, KNNI, and (ABE-FCMI)) on Desharnais dataset with MNAR missingness mechanism and different missing ratio (Dish-Miss1,Dish-Miss2). First the experimental results for each incomplete dataset is evaluated individually, second a comparison between imputation methods is evaluated based on all given incomplete datasets.

A. Effects of MI, KNNI and ABE-FCMI on Desharnais Dataset

As discussed before Desharnais dataset contain missing values. In projects number 38, 44, the TeamExp feature values are missing. In projects number 38, 66, and 75, the ManagerExp feature values are missing. It can be concluded that Desharnais dataset have relatively lower number of missing values compared to other given incomplete datasets in this study. In step 1 Desharnais dataset is taken as incomplete dataset. In step 2 missing data imputation is performed using MI, KNNI, and (ABE-FCMI). In step 3 accuracy evaluation of ABE is measured for each imputation technique. Three-fold cross validation technique has been used to generate the results. The overall empirical process can be seen in Fig. 5. Table II shows MMRE results of imputation methods on ABE, while Table III shows the PRED(25) results of imputation methods on ABE, and Table IV shows SA results of imputation methods.

As seen in Table II, MI and (ABE-FCMI) achieved the lowest value of MMRE as 0.02622 and 0.02631 respectively with regard to the average of three folds. It is followed by KNNI where the value of MMRE is 0.02651. It is observed that the lowest value of MMRE is achieved by MI due to lower number of missing data in Desharnais dataset. Table III shows the PRED (0.25) results obtained from applying imputation methods to Desharnais dataset based on three-fold cross validation. As can be seen the PRED values are the same for all imputation methods. The SA results for imputation methods are given in Table IV. MI and (ABE-FCMI) achieved best SA results with values 56.66670, 56.49223 respectively, while KNNI achieved 56.38617 value for SA accuracy measure. It is

observed that the best value of SA is achieved by MI due to lower number of missing data in Desharnais dataset.

TABLE III. MMRE RESULTS OF IMPUTATION METHODS ON ABE MODEL FOR DESHARNAIS DATASET

Imputation Method	FOLD1	FOLD2	FOLD3	Average
Mean	0.01953	0.03234	0.02678	0.02622
KNN	0.02023	0.03266	0.02665	0.02651
ABE-FCMI	0.01979	0.03238	0.02672	**0.02631**

TABLE IV. PRED (0.25) RESULTS OF IMPUTATION METHODS ON ABE MODEL FOR DESHARNAIS DATASET

Imputation Method	FOLD1	FOLD2	FOLD3	Average
Mean	33.33333	40.74074	37.03704	37.03704
KNN	33.33333	40.74074	37.03704	37.03704
ABE-FCMI	33.33333	40.74074	37.03704	37.03704

TABLE V. SA RESULTS OF IMPUTATION METHODS ON ABE MODEL FOR DESHARNAIS DATASET

Imputation Method	FOLD1	FOLD2	FOLD3	Average
Mean	60.00657	50.91008	59.08344	56.66670
KNN	58.77629	50.69166	59.69057	56.38617
ABE-FCMI	59.38786	50.80258	59.28626	56.49223

B. Effects of MI, KNNI and ABE-FCMI on Desh-Miss1 Dataset

As discussed before Desh-Miss1 dataset have 28.395% missing row ratio (23 out of 81 projects have missing values) and 3.33 % missing cell ratio (24 missing cells out of 720 cells) with MNAR missingness mechanism. Desh-Miss1 dataset is incomplete dataset generated from Desharnais dataset.

As can be seen from Table V, (ABE-FCMI) achieved lower MMRE among all other imputation methods on ABE model with value (0.02589). It is followed by KNNI and MI with values 0.02608, 0.02634, respectively. (ABE-FCMI) archived higher PRED with value 38.27160 as given from Table VI. It is followed by KNNI and MI with the same value 35.80247. Best SA value is achieved by (ABE-FCMI) with value 56.97777 as observed from Table VII. The calculated SA values for KNNI, MI were 56.39966, 55.93544 respectively. As a result (ABE-FCMI) accomplished significant improvement compared to KNNI and MI on the selected accuracy evaluation measures (MMRE, PRED, and SA) for ABE estimation model applied for Desh-Miss1 incomplete dataset.

C. Effects of MI, KNNI and ABE-FCMI on Desh-Miss2 Dataset

As discussed before Desh-Miss2 dataset have 69.135 % missing row ratio (56 out of 81 projects have missing values) 7.916 % missing cell ratio (57 missing cells out of 720 cells) with MNAR missingness mechanism. Desh-Miss2 dataset is incomplete dataset generated from Desharnais dataset. As can be seen from Table VIII, ABE-FCMI achieved lower MMRE among all other imputation methods on ABE model with value (0.02557). It is followed by KNNI and MI with values 0.02693, 0.02794 respectively. The highest PRED values for all applying imputation methods on ABE for Desh-Miss2 dataset was achieved by (ABE-FCMI) with value 43.20988 as given from Table IX. It is followed by KNNI and MI with the same value 38.2716.

The SA results for ABE model on Desh-Missing2 after applying the selected imputation methods are given in Table X. ABE-FCMI accomplished best result for SA measure with value 56.92689. It is followed by KNNI and MI with values 56.80289, 55.80017 respectively. As a result, ABE-FCMI accomplished significant improvement compared to KNNI and MI on the selected accuracy evaluation measures (MMRE, PRED, and SA) for ABE estimation model applied for Desh-Miss2 incomplete dataset.

TABLE VI. MMRE RESULTS OF IMPUTATION METHODS ON ABE MODEL FOR DESH-MISS1 DATASET

Imputation Method	FOLD1	FOLD2	FOLD3	Average
Mean	0.01954	0.0322	0.02728	0.02634
KNN	0.01935	0.03161	0.02728	0.02608
(ABE-FCMI)	0.01899	0.03225	0.02642	0.02589

TABLE VII. PRED (25) RESULTS OF IMPUTATION METHODS ON ABE MODEL FOR DESH-MISS1 DATASET

Imputation Method	FOLD1	FOLD2	FOLD3	Average
Mean	29.62963	44.44444	33.33333	35.80247
KNN	33.33333	37.03704	37.03704	35.80247
(ABE-FCMI)	33.33333	40.74074	40.74074	38.27160

TABLE VIII. SA RESULTS OF IMPUTATION METHODS ON ABE MODEL FOR DESH-MISS1 DATASET

Imputation Method	FOLD1	FOLD2	FOLD3	Average
Mean	59.04778	50.47609	58.28246	55.93544
KNN	60.2602	48.96933	59.96946	56.39966
(ABE-FCMI)	60.06159	50.91122	59.96049	56.97777

TABLE IX. MMRE RESULTS OF IMPUTATION METHODS ON ABE MODEL FOR DESH-MISS2 DATASET

Imputation Method	FOLD1	FOLD2	FOLD3	Average
Mean	0.01908	0.03284	0.0319	0.02794
KNN	0.01887	0.03407	0.02785	0.02693
(ABE-FCMI)	0.018	0.03055	0.02816	0.02557

TABLE X. PRED (25) RESULTS OF IMPUTATION METHODS ON ABE MODEL FOR DESH-MISS2 DATASET

Imputation Method	FOLD1	FOLD2	FOLD3	Average
Mean	37.03704	33.33333	44.44444	38.2716
KNN	40.74074	29.62963	44.44444	38.2716
(ABE-FCMI)	51.85185	29.62963	48.14815	43.20988

D. Comparison of Imputation Methods for ABE

A comparison between all selected imputation methods (MI, KNNI, and (ABE-FCMI) on all selected incomplete datasets (Desharnais, Dish-Miss1, and Dish-Miss2) for ABE estimating model is presented in Table XI. As the percentages of missing values are increased the calculated MMRE values for imputation methods are generally increased as shown in Table XII. For example MMRE values for MI are increased sequentially (0.02622, 0.02634, and 0.02794) for Desharnais, Desh-Miss1, and Desh-Miss2 incomplete datasets. Fig. 6 shows comparison based on MMRE values for MI, KNNI, and (ABE-FCMI) applied for ABE estimation model for all selected incomplete dataset in this study.it is observed that the MMRE values are increased as the number of missing values for incomplete datasets (Desharnais,Dish-Miss1,Desh-Miss2) are grown also.

As can be seen from Table XI, PRED values for MI and KNNI imputation methods have equal values as the percentage of missing data are increased. For example in Desharnias dataset PRED values for MI and KNNI are 37.03704. With increased number of missing values from Dish-Miss1 to Dish-Miss2 datasets, the PRED values for MI and KNNI are equal (35.80247) in Dish-Miss1 dataset, and also for Dish-Miss2 dataset with PRED value (38.2716) for MI and KNNI. Fig. 7 shows comparison based on PRED values for MI, KNNI, and (ABE-FCMI) applied for ABE estimation model for all selected incomplete dataset in this study.it is observed that (ABE-FCMI) improved significantly PRED values measure for Dish-Miss1 and Dish-Miss2 datasets with values 38.2716, 43.20988 respectively. It can be seen that (ABE-FCMI) successfully improve PRED measure although with increased number of missing values. MMRE and PRED are considered as biased accuracy measurements in ABE model and produced asymmetric distribution, there is a need for unbiased accuracy evaluation using SA measure [53-55]. A SA evaluation criterion is applied in this study for ABE estimation model. As can be seen from Table XI, the SA values are decreased as the numbers of missing values are increased from Desh-Miss1 to Desh-Miss2 incomplete datasets. For example the SA values for MI are 55.93544, 55.80017 respectively for Desh-Miss1 and Dish-Miss2. Another example the SA values for (ABE-FCMI) are 56.97777, 56.92689 respectively for Desh-Miss1 and Dish-Miss2.

Fig. 8 shows comparison based on SA values for MI, KNNI, and (ABE-FCMI) applied for ABE estimation model for all selected incomplete dataset in this study. As can be seen that the SA values are decreased as the number of missing values are increased, (ABE-FCMI) achieved the highest SA values in Desh-Miss1 and Dish-Miss2 with values 56.97777, 56.92689 respectively. For Desharnais dataset due to lower number of missing values (4 missing rows, 5 missing cells) compared to other incomplete datasets (Desh-Miss1, Desh-Miss2), (ABE-FCMI) achieved second highest SA value (56.49223). As a result (ABE-FCMI) achieved best results of the performance accuracy measures (MMRE, PRED, and SA) compared to MI and KNNI for ABE estimation model in incomplete datasets (Dish-Miss1, Dish-Miss2). Due to low number of missing cases in Desharnais dataset (ABE-FCMI) achieved second winner after MI method. The effectiveness of (ABE-FCMI) method to improve ABE accuracy result for Desharnais dataset is proven through the experimental part of this study. (ABE-FCMI) imputes missing datasets with more realistic values compared to MI and KNNI.

TABLE XI. SA RESULTS OF IMPUTATION METHODS ON ABE MODEL FOR DESH-MISS2 DATASET

Imputation Method	FOLD1	FOLD2	FOLD3	Average
Mean	60.36604	49.92606	57.10841	55.80017
KNN	61.1994	49.62232	59.58695	56.80289
(ABE-FCMI)	62.36436	50.74684	57.66948	56.92689

TABLE XII. COMPARISION OF (MI, KNNI, AND (ABE-FCMI)) IMPUTATION METHODS FOR (DESHARNAIS, DESH-MISS1, AND DESH-MISS2) FOR ABE MODEL

DATASET	MI			KNNI			(ABE-FCMI)		
	MMRE	PRED(25)	SA	MMRE	PRED(25)	SA	MMRE	PRED(25)	SA
1	0.02622	37.03704	56.66670	0.02651	37.03704	56.38617	0.02631	37.03704	56.49223
2	0.02634	35.80247	55.93544	0.02608	35.80247	56.39966	0.02589	38.2716	56.97777
3	0.02794	38.2716	55.80017	0.02693	38.2716	56.80289	0.02557	43.20988	56.92689

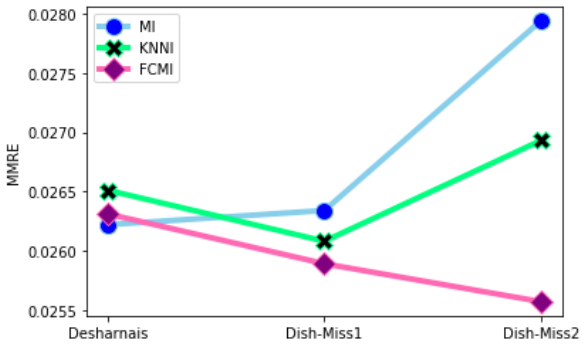

Fig. 6. Comparison of MMRE of (MI, KNNI, (ABE-FCMI)) for ABE Model.

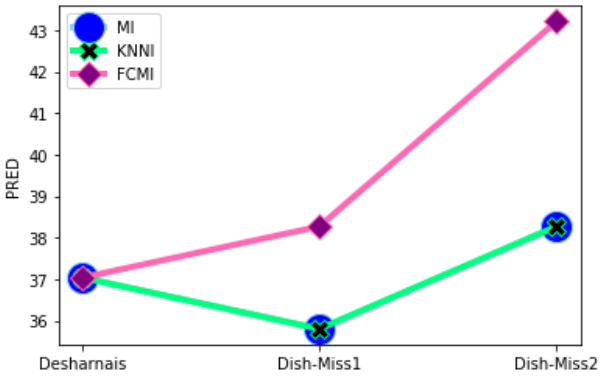

Fig. 7. Comparison of PRED (25) of (MI, KNNI, (ABE-FCMI)) for ABE Model.

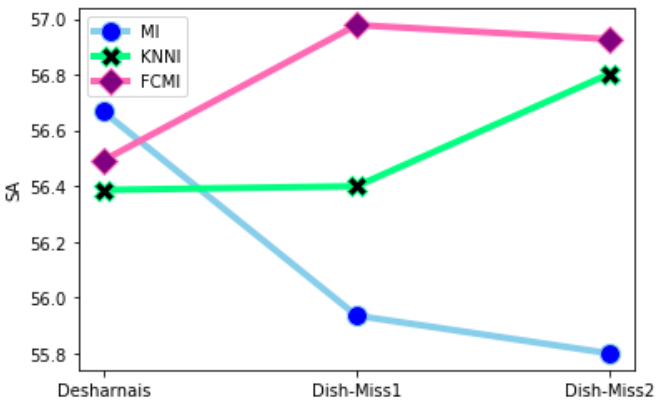

Fig. 8. Comparison of SA of (MI, KNNI, (ABE-FCMI)) for ABE Model.

VIII. Threats to Validity

In this empirical study, an evaluation of three imputation techniques using MNAR missingness mechanism and different MD percentages has been reported. It is difficult to carry out all possible scenarios, so some limitation may exist in this study.

A. Internal Validity

Internal validity is concerned with threats related to the scope of the study. In this study, an investigation attempted to simulate scenarios with MNAR missingness mechanism as well as different MD percentages. Generation of MD process for MNAR mechanism might considered as internal thread. A random selection of attribute for MD generation in the studied dataset is used. In this study we simulate tow incomplete datasets with different MD percentages; a threat might come from MD percentages as well as we investigate only MNAR mechanism.

B. External Validity

External validity is related to threats that are concerned with empirical design and result generalization. In this experimental study, we investigate Desharnais dataset as one of the most common datasets in the field of software effort estimation. Recent research studies investigate Desharnais dataset imputation for ABE performance evaluation [39, 42, 44]. Desharnais dataset is considered relatively small with 81 software projects only, and contained only numerical attributes, these might be considered as external threats, Table XIII.

IX. Conclusion and Future Work

The quality of the dataset plays a vital role for accurate software effort estimation process. Handling missing data problem is a major challenge to increase the quality of the dataset used for effort prediction. ABE as wide accepted effort estimation model depend mainly on the completed historically dataset for effort prediction, therefore confronting missing values in previously completed projects will improve the accuracy of ABE prediction. Different missing data imputation techniques have been used for ABE model including MI and KNNI. MI method is considered as static imputation without analyzed the dynamic nature for each missing case in the feature concerned in the incomplete software project. KNNI used Euclidian similarity measure to whole completed dataset to identify similar donor cases which may or not be related to the incomplete software project. In this study an imputation technique based on FCM clustering have been proposed for ABE model. The proposed (ABE-FCMI) technique is investigated for Desharnias dataset with different missing ratio and MNAR missingness mechanism. Experimental results suggest that ABE model using FCM imputation have provided significant improvement against ABE model using either MI or KNNI imputation methods. ABE Performance improvement of the proposed imputation method is based that FCM algorithm clustered software projects into homogeneous clusters based on the selected dataset attributes. Based on the completed dataset FCM algorithm identifies cluster centers. Imputation values for each incomplete project is calculated based on their distance and membership to the cluster centers identified before. (ABE-FCMI) identifies more reliable donors cases to the incomplete software project to impute missing values compared to KNNI and MI.

The Performance of ABE model has been positively affected with MD imputation techniques used in this study for incompleted datasets as seen in accuracy results. In comparison, (ABE-FCMI) significantly outperforms MI and KNNI in missing data imputation for ABE model in Desh-Miss1 and Desh-Miss2 incomplete datasets. For Desharnais dataset due to low number of missing values, there is no significant difference between the three imputations techniques used in Desharnais dataset. The fuzzy clustering nature of (ABE-FCMI) to identify groups of most similar projects indicate that it imputes more reliable values compared to MI and slightly better than KNNI on small datasets.

The study results have shown that as the percentage of missing data of MNAR mechanism increased from Desh-Miss1 to Desh-Miss2 incomplete dataset, the accuracy of ABE model is decreased using MI and KNNI imputation methods, however (ABE-FCMI) improved ABE accuracy although with increased percentage of missing data of MNAR mechanism.

The investigated software engineering dataset in this study is relatively small with 81 software projects only. We suggested investigating (ABE-FCMI) for large software engineering datasets to generalize our results. Numerical missing value imputation is the focus of this study; mixed (numerical and categorical) missing data imputation is required to verify the performance of (ABE-FCMI) method for ABE model.

References

[1] Jones, C., Estimating Software Costs: Bringing Realism to Estimating 2007. Tata McGraw-Hill.

[2] Wen, J., et al., Systematic literature review of machine learning based software development effort estimation models. Information and Software Technology, 2012. 54(1): p. 41-59.

[3] Jorgensen, M. and M. Shepperd, A systematic review of software development cost estimation studies. IEEE Transactions on software engineering, 2006. 33(1): p. 33-53.

[4] Idri, A., F. azzahra Amazal, and A. Abran, Analogy-based software development effort estimation: A systematic mapping and review. Information and Software Technology, 2015. 58: p. 206-230.

[5] Strike, K., K. El Emam, and N. Madhavji, Software cost estimation with incomplete data. IEEE Transactions on Software Engineering, 2001. 27(10): p. 890-908.

[6] Huang, J., Y.-F. Li, and M. Xie, An empirical analysis of data preprocessing for machine learning-based software cost estimation. Information and software Technology, 2015. 67: p. 108-127.

[7] Song, Q., et al., Can k-NN imputation improve the performance of C4. 5 with small software project data sets? A comparative evaluation. Journal of Systems and software, 2008. 81(12): p. 2361-2370.

[8] Idri, A., I. Abnane, and A. Abran. Systematic mapping study of missing values techniques in software engineering data. in 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). 2015. IEEE.

[9] Almutlaq, A.J.H. and D.N. Jawawi. Missing Data Imputation Techniques for Software Effort Estimation: A Study of Recent Issues and Challenges. in International Conference of Reliable Information and Communication Technology. 2019. Springer.

[10] Myrtveit, I., E. Stensrud, and U.H. Olsson, Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods. IEEE Transactions on Software Engineering, 2001. 27(11): p. 999-1013.

[11] Lin, W.-C. and C.-F. Tsai, Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review, 2020. 53(2): p. 1487-1509.

[12] Huang, J., et al., Cross-validation based K nearest neighbor imputation for software quality datasets: An empirical study. Journal of Systems and Software, 2017. 132: p. 226-252.

[13] Shepperd, M. and C. Schofield, Estimating software project effort using analogies. IEEE Transactions on software engineering, 1997. 23(11): p. 736-743.

[14] Kadoda, G., et al. Experiences using case-based reasoning to predict software project effort. in Proceedings of the EASE 2000 conference, Keele, UK. 2000. Citeseer.

[15] Walkerden, F. and R. Jeffery, An empirical study of analogy-based software effort estimation. Empirical software engineering, 1999. 4(2): p. 135-158.

[16] Angelis, L. and I. Stamelos, A simulation tool for efficient analogy based cost estimation. Empirical software engineering, 2000. 5(1): p. 35-68.

[17] Little, R.J. and D.B. Rubin, The analysis of social science data with missing values. Sociological Methods & Research, 1989. 18(2-3): p. 292-326.

[18] Li, J., A. Al-Emran, and G. Ruhe. Impact analysis of missing values on the prediction accuracy of analogy-based software effort estimation method AQUA. in First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007). 2007. IEEE.

[19] Cartwright, M.H., M.J. Shepperd, and Q. Song. Dealing with missing software project data. in Proceedings. 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry (IEEE Cat. No. 03EX717). 2004. IEEE.

[20] Twala, B. and M. Cartwright. Ensemble imputation methods for missing software engineering data. in 11th IEEE International Software Metrics Symposium (METRICS'05). 2005. IEEE.

[21] Fujikawa, Y. and T. Ho. Cluster-based algorithms for dealing with missing values. in Pacific-Asia conference on knowledge discovery and data mining. 2002. Springer.

[22] Banerjee, A., et al., Clustering with Bregman divergences. Journal of machine learning research, 2005. 6(10).

[23] Patil, B.M., R.C. Joshi, and D. Toshniwal. Missing value imputation based on k-mean clustering with weighted distance. in International Conference on Contemporary Computing. 2010. Springer.

[24] Zhang, S., et al., Missing value imputation based on data clustering, in Transactions on computational science I. 2008, Springer. p. 128-138.

[25] Luengo, J., S. García, and F. Herrera, On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowledge and information systems, 2012. 32(1): p. 77-108.

[26] Sefidian, A.M. and N. Daneshpour, Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model. Expert Systems with Applications, 2019. 115: p. 68-94.

[27] Aydilek, I.B. and A. Arslan, A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Information Sciences, 2013. 233: p. 25-35.

[28] Rahman, M.G. and M.Z. Islam, Missing value imputation using a fuzzy clustering-based EM approach. Knowledge and Information Systems, 2016. 46(2): p. 389-422.

[29] Carneiro, C., et al., Advanced data mining method for discovering regions and trajectories of moving objects:"ciconia ciconia" scenario, in The European Information Society. 2008, Springer. p. 201-224.

[30] García, S., J. Luengo, and F. Herrera, Data preprocessing in data mining. Vol. 72. 2015: Springer.

[31] Li, D., et al. Towards missing data imputation: a study of fuzzy k-means clustering method. in International conference on rough sets and current trends in computing. 2004. Springer.

[32] Zadeh, L.A., Fuzzy sets. Information and control, 1965. 8(3): p. 338-353.

[33] Di Nuovo, A.G., Missing data analysis with fuzzy C-Means: A study of its application in a psychological scenario. Expert Systems with Applications, 2011. 38(6): p. 6793-6797.

[34] Timm, H., C. Döring, and R. Kruse, Different approaches to fuzzy clustering of incomplete datasets. International Journal of Approximate Reasoning, 2004. 35(3): p. 239-249.

[35] Bezdek, J.C., R. Ehrlich, and W. Full, FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 1984. 10(2-3): p. 191-203.

[36] Dunn, J.C., Well-separated clusters and optimal fuzzy partitions. Journal of cybernetics, 1974. 4(1): p. 95-104.

[37] Hathaway, R.J. and J.C. Bezdek, Fuzzy c-means clustering of incomplete data. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2001. 31(5): p. 735-744.

[38] Sentas, P. and L. Angelis, Categorical missing data imputation for software cost estimation by multinomial logistic regression. Journal of Systems and Software, 2006. 79(3): p. 404-414.

[39] Idri, A., I. Abnane, and A. Abran, Missing data techniques in analogy-based software development effort estimation. Journal of Systems and Software, 2016. 117: p. 595-611.

[40] Abnane, I. and A. Idri. Evaluating fuzzy analogy on incomplete software projects data. in 2016 IEEE Symposium Series on Computational Intelligence (SSCI). 2016. IEEE.

[41] Huang, J., et al. An empirical analysis of three-stage data-preprocessing for analogy-based software effort estimation on the ISBSG data. in 2017 IEEE International Conference on Software Quality, Reliability and Security (QRS). 2017. IEEE.

[42] Idri, A., I. Abnane, and A. Abran, Support vector regression ‐ based imputation in analogy ‐ based software development effort estimation. Journal of Software: Evolution and Process, 2018. 30(12): p. e2114.

[43] Abnane, I. and A. Idri. Improved analogy-based effort estimation with incomplete mixed data. in 2018 Federated Conference on Computer Science and Information Systems (FedCSIS). 2018. IEEE.

[44] Muhammad Arif Shah, D.N.A.J., Mohd Adham Isa, Karzan Wakil, Muhammad Younas, Ahmed Mustafa, MINN: A Missing Data Imputation Technique for Analogy-based Effort Estimation. International Journal of Advanced Computer Science and Applications, 2019. 10(2).

[45] Abnane, I., et al. Analogy software effort estimation using ensemble KNN imputation. in 2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA). 2019. IEEE.

[46] Abnane, I., A. Idri, and A. Abran, Fuzzy case ‐ based ‐ reasoning ‐ based imputation for incomplete data in software engineering repositories. Journal of Software: Evolution and Process, 2020: p. e2260.

[47] Horton, N.J. and K.P. Kleinman, Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. The American Statistician, 2007. 61(1): p. 79-90.

[48] Mockus, A., Missing data in software engineering, in Guide to advanced empirical software engineering. 2008, Springer. p. 185-200.

[49] Beretta, L. and A. Santaniello, Nearest neighbor imputation algorithms: a critical evaluation. BMC medical informatics and decision making, 2016. 16(3): p. 74.

[50] Zhang, S., Nearest neighbor selection for iteratively kNN imputation. Journal of Systems and Software, 2012. 85(11): p. 2541-2552.

[51] Desharnais, J., Analyse statistique de la productivitie des projects informatique a partie de la technique des point des function. Masters Thesis University of Montreal, 1989.

[52] Idri, A., I. Abnane, and A. Abran, Evaluating Pred (p) and standardized accuracy criteria in software development effort estimation. Journal of Software: Evolution and Process, 2018. 30(4): p. e1925.

[53] Shepperd, M. and S. MacDonell, Evaluating prediction systems in software project estimation. Information and Software Technology, 2012. 54(8): p. 820-827.

[54] Foss, T., et al., A simulation study of the model evaluation criterion MMRE. IEEE transactions on software engineering, 2003. 29(11): p. 985-995.

[55] Myrtveit, I., E. Stensrud, and M. Shepperd, Reliability and validity in comparative studies of software prediction models. IEEE Transactions on Software Engineering, 2005. 31(5): p. 380-391.

APPENDIX

TABLE XIII. SAMPLE DATA FROM ORIGINAL DESHARNAIS DATASET

TeamExp	ManagerExp	Length	Transactions	Entities	PointsAdjust	Envergure	PointsNonAjust	Effort
2.0	1.0	9.0	119.0	42.0	161.0	25.0	145.0	2569.0
1.0	2.0	**13.0**	186.0	52.0	238.0	25.0	214.0	3913.0
3.0	1.0	12.0	172.0	88.0	**260.0**	30.0	247.0	7854.0
3.0	4.0	4.0	78.0	38.0	116.0	24.0	103.0	2422.0
4.0	1.0	21.0	167.0	**99.0**	266.0	24.0	237.0	4067.0
2.0	**1.0**	17.0	146.0	112.0	258.0	40.0	271.0	9051.0

TABLE XIV. SAMPLE DATA OF INCOMPLETE DESHARNAIS DATASET (DESH-MISS2) OF STEP 1 USING MNAR MECHANISM WITH 69.135 % OF MD, WHERE NULL DENOTES THE REMOVED DATA

TeamExp	ManagerExp	Length	Transactions	Entities	PointsAdjust	Envergure	PointsNonAjust	Effort
NULL	1.0	9.0	119.0	42.0	161.0	25.0	145.0	2569.0
1.0	2.0	**NULL**	186.0	52.0	238.0	25.0	214.0	3913.0
3.0	1.0	12.0	172.0	88.0	**NULL**	30.0	247.0	7854.0
3.0	4.0	4.0	78.0	38.0	116.0	24.0	103.0	2422.0
4.0	1.0	21.0	167.0	**NULL**	266.0	24.0	237.0	4067.0
2.0	**NULL**	17.0	146.0	112.0	258.0	40.0	271.0	9051.0

TABLE XV. SAMPLE DATA OF (DESH-MISS2) OF STEP 2 IMPUTED USING (FCMI-ABE) IMPUTATION UNDER MNAR MECHANISM WITH 69.135 % OF MD. IMPUTED VALUES ARE INDICATED IN BOLD

TeamExp	ManagerExp	Length	Transactions	Entities	PointsAdjust	Envergure	PointsNonAjust	Effort
2.315	1.0	9.0	118.999	42.0	161.0	25.0	145.0	2569.0
1.0	2.0	**8.372**	186.0	52.0	238.0	25.0	214.0	3913.0
3.0	1.0	12.0	172.0	88.0	**217.154**	30.0	246.999	7854.0
3.0	4.0	4.0	78.0	38.0	116.0	24.0	103.0	2422.0
4.0	1.0	21.0	167.0	**91.639**	266.0	24.0	236.999	4067.0
2.0	**2.497**	17.0	146.0	112.0	258.0	40.0	270.999	9051.0