EXPERT SYSTEMS WITH APPLICATIONS 00 (2020) 000–000

HOSTED BY
ELSEVIER

SciVerse ScienceDirect

# Evidential Reasoning for Preprocessing Uncertain Categorical Data for Trustworthy Decisions: An Application on Healthcare and Finance

***Swati Sachan (Corresponding author)***
*Post-doctoral Research Associate*
Decision and Cognitive Science Research Centre (DCSRC)
The University of Manchester
Booth Street West
Manchester, M15 6PB
United Kingdom
swati.sachan@manchester.ac.uk

***Fatima Almaghrabi***
*PhD Researcher*
Decision and Cognitive Science Research Centre (DCSRC)
The University of Manchester
Booth Street West
Manchester, M15 6PB
United Kingdom
fatima.almaghrabi@postgrad.manchester.ac.uk

***Jian-Bo Yang***
*Director of Decision and Cognitive Sciences Research Centre (DCS)*
Decision and Cognitive Science Research Centre (DCSRC)
The University of Manchester
Booth Street West
Manchester, M15 6PB
United Kingdom
jian-bo.yang@manchester.ac.uk

***Dong-Ling Xu***
*Chair Professor of Decision Science and Systems*
Decision and Cognitive Science Research Centre (DCSRC)
The University of Manchester
Booth Street West
Manchester, M15 6PB
United Kingdom
Ling.Xu@manchester.ac.uk

*Abstract* — **The uncertainty attributed by discrepant data in AI-enabled decisions is a critical challenge in highly regulated domains such as health care and finance. Ambiguity and incompleteness due to missing values in output and input attributes, respectively, is ubiquitous in these domains. It could have an adverse impact on a certain unrepresented set of people in the training data without a developer's intention to discriminate. The inherently non-numerical nature of categorical attributes than numerical attributes and the presence of incomplete and ambiguous categorical attributes in a dataset increases the uncertainty in decision-making. This paper addresses the challenges in handling categorical attributes as it is not addressed comprehensively in previous research. Three sources of uncertainties in categorical attributes are recognised in this research. The informational uncertainty, unforeseeable uncertainty in the decision task environment, and the uncertainty due to lack of pre-modelling explainability in categorical attributes are addressed in the proposed methodology on maximum likelihood evidential reasoning (MAKER). It can transform and impute incomplete and ambiguous categorical attributes into interpretable numerical features. It utilises a notion of weight and reliability to include subjective expert preference over a piece of evidence and the quality of evidence in a categorical attribute, respectively. The MAKER framework strives to integrate the recognised uncertainties in the transformed input data that allow a model to perceive data limitations during the training regime and acknowledge doubtful predictions by supporting trustworthy pre-modelling and post modelling explainability. The ability to handle uncertainty and its impact on explainability is demonstrated on a real-world healthcare and finance data for different missing data scenarios in three types of AI algorithms: deep-learning, tree-based, and rule-based model.**

**Keywords—categorical, uncertainty, decision-making, evidential reasoning, trustworthy**

## 1. Introduction

Data preprocessing is a crucial step towards the achievement of high-quality data. The perfection in the sense of completeness, lack of ambiguity, meaningfulness, and correctness (Wand & Wang, 1996) in high-quality data promote trustworthy and reliable insights from data-driven decision-making systems. The preprocessing of categorical attributes is much more challenging than numerical attributes in a dataset. Missing data imputation techniques and data-driven models can process data only in a numerical format. Therefore, categorical attributes are first transformed into numbers. The techniques applied for preprocessing the input data can have a significant impact on model performance and explainability. Primarily, the explainability of a data-driven autonomous system has two stages: pre-modelling explainability to gain an understanding of the data imputed in a decision-making system and post-modelling to understand the reasoning behind a decision generated by a system (Sachan, Yang, & Xu, 2020). Pre-modelling explainability is in demand due to a rise in ethical concerns regarding the justification and legitimacy of the data used in the model.

The challenge to overcome uncertainty is critical in a decision-making system enabled by artificial intelligence or machine learning (AI/ML) techniques. The presence of uncertainty is ubiquitous in a realistic setting and is critical in highly regulated domains, such as finance, insurance, healthcare, and medical science. All uncertainties cannot be eliminated; however, it cannot be fully ignored. It constrains the sustainability of AI/ML decision-making applications (Walker, Haasnoot, & Kwakkel, 2013). Three sources of uncertainties - incomplete information, inadequate understanding, and undifferentiated alternatives in decision-making were identified by Lipshitz & Strauss (Lipshitz & Strauss, 1997). A review on informational, environmental, and intentional uncertainty in AI-enabled decision-making was conducted by Wu and Shang (Wu & Shang, 2020). The uncertainty is mostly associated with incomplete information (Milliken, 1987). In AI-enabled algorithms, the incompleteness of information points to missing values and the lack of sufficient instances in a dataset. A paper has presented an adaptive algorithm to find an optimal imputation method for an AI/ML classifier based on missing data characteristics (Sim, Kwon, & Lee, 2016). Previous research in different areas has investigated techniques to treat missing data, such as healthcare (Masconi, Matsha, Echouffo-Tcheugui, Erasmus, & Kengne, 2015), financial credit scoring (Lan, Xu, Ma, & Li, 2020), customer satisfaction (Maddulapallia, Yang, & Xu, 2012), diagnostic and prognostic (Razavi-Far, Chakrabarti, Saif, & Zio, 2019) and psychology (Roth, 1994) and for multiple types of data, for instance, survey data (Wang & Wang, 2009) and longitudinal data (Huque, Carlin, Simpson, & Lee, 2018). The presence of categorical attributes increases the information uncertainty (Qin, Xia, & Prabhakar, 2011). Dealing with numerical data is often easier than categorical data as the semantics pertaining to each categorical value in a categorical attribute does not require transformation to a numerical format. The uncertainties incurred due to incompleteness and ambiguity in categorical attributes have not been addressed adequately in previous research. It is valuable to develop techniques to handle uncertain categorical attributes that reduce vulnerability towards a plausible future which demands trustworthy decisions.

The categorical attributes in real-world data are usually missing and have inherently non-numerical nature, which present challenges in the development of AI/ML models. A dataset can have three types of missing and non-missing patterns: complete, incomplete, and ambiguous (Sachan S. , Yang, Xu, Benavides, & Li, 2020). A complete dataset has full records of all input and output attributes. An incomplete dataset has missing records in input attributes. An ambiguous dataset has missing records in an output attribute. Data can be called unambiguous if all outputs are available. A dataset is incomplete and ambiguous if it has missing values in both input and output attributes. The missing and non-missing patterns are shown in Figure 1. The incomplete dataset (only input attributes missing) can have three types of missing mechanisms: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Rubin, 1976). The missing mechanism by MCAR is independent of observed and unobserved data, whereas the missing mechanism by MAR is independent of unobserved data and dependent on observed data. The missing mechanism by MNAR is only dependent on unobserved data. The sensitivity analysis to assess the robustness of the assumptions for multiple imputation techniques for MCAR, MAR, and MNAR types of missing data is demonstrated in (Sidi & Harel, 2018). The case-wise deletion, also known as the list-wise deletion (Briggs, Clark, Wolstenholme, & Clarke, 2003), and the imputation of missing values are two conventional methods to handle missing values. The case-wise deletion utilizes samples or cases with no missing values in all the attributes of a dataset. The practice of complete-case analysis could result in a significant loss of information (Jamshidian & Mata, 2007). The missing values imputation method estimates a set of plausible values for the missing values using the distribution of the observed data. The exclusion of incomplete cases by a deletion method is not always possible for most use cases in different domains as it could result in biased inference due to poor representation of the entire data by complete cases (Brown, 1994) (Graham, Hofer, & MacKinnon, 1996). For instance, electoral registration plays a vital role in credit evaluation and address checks for loan and credit card applications in the United Kingdom. However, customers cannot be turned down if their electoral information is unavailable for their current and previous addresses in historical credit data. Similarly, symptoms of diseases or conditions and the medical history of patients are recorded based on the list of questionnaires that consist of multiple-choice options and Likert scale questions. Such dataset could gather incomplete information when specific questions do not apply to a respondent. This may be because the respondent does not want to disclose private information or is merely uncertain and incapable of answering the questions. Such circumstances present a challenge to leverage AI/ML to foster domain expertise because there will always be a situation where data is incomplete and ambiguous. Most AI/ML models cannot work with missing values in the datasets. Various missing data imputation techniques are developed to handle this issue. The existing missing data imputation techniques are discussed in Section 2, literature review. The basic idea is to replace the missing values with the predicted values obtained from the observation data.

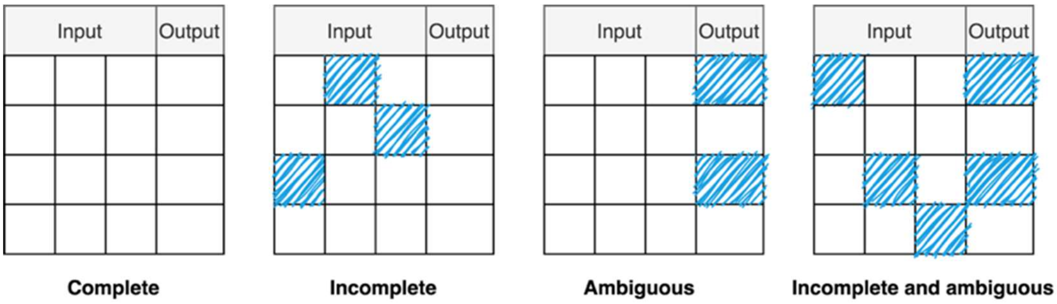


Figure 1. Missing and non-missing pattern

## 2. Literature Review

### 2.1 Data imputation and transformation techniques

An instance contains a piece of evidence for each attribute. A decision-making system provides a decision for an instance by analysing the pieces of evidence. In a tabular dataset, a row represents an instance, and a column represents an attribute. The issue of missing values in all instances must be resolved before sending to an automated decision-making system. The imputed missing value should be meaningful for subsequent analysis by the system. Most missing value imputation techniques resolve this task by analysing only complete cases. The complete cases are obtained by removing instances with one or more missing values of attributes. However, discarding incomplete information may result in loss of information, and it could introduce biases if missing data is not random. The quantification of the loss of information was demonstrated by Richman, Trafalis, and Adrianto (2009) in an example where a dataset with 100 instances had 1% chance of missing the values in each attribute independently. The expected proportion of complete cases would be $0.99^{100} = 0.366$, which retains only $^{.366}/_{.99} \cong 36.9\%$ of the data. The loss of information decreases the accuracy due to an increase in variance in the data.

Most prevalent techniques for the imputation of missing values are Multivariate Imputation by Chained Equations (MICE), expectation-maximization (EM) algorithm, K-nearest neighbour (KNN) algorithm, missForest based on random forest algorithm, and other supervised machine learning algorithms, such as an artificial neural network (ANN) and support vector machine (SVM). In MICE, the initial values in attributes are imputed simply by its mean or frequency. They are then replaced by values predicted by a linear regression model between a dependent attribute with missing values and independent attributes with complete cases (Azur, Stuart, Frangakis, & Leaf, 2011). The MICE runs a series of regression models for each attribute with missing values and works under the assumption that data is missing due to the MAR and MCAR mechanism. The regularised regression model can be used to minimise the loss of function by imposing some penalties. This technique is called MICE by regularised regression. The superiority of different regularised regression models, such as lasso, elastic net, and adaptive lasso in terms of biases in imputed missing values in high-dimensional data is presented in (Deng, Chang, Ido, & Long, 2016). The expectation-maximisation (EM) algorithm works under the assumption that the model's parameters are known (Briggs, Clark, Wolstenholme, & Clarke, 2003). It initially fills missing values with the best guess under the initial estimate of the unknown parameters. The EM algorithm iterates in two cycles: the expectation step (E-step) and the maximisation step (M-step). In E-step, missing values are estimated by log-likelihood of complete data with respect to incomplete data or missing values. The model parameters are maximised in M-step by utilising data updated in E-step (Dempster, Laird, & Rubin, 1977). The parameters are re-estimated from the updated complete dataset. The EM algorithm was proposed as an iterative regression approach to converge with an appropriate imputation value (Meng & Rubin, 1991). The missing values imputed by iterative regression-based methods can be explained by coefficients (parameters) of the independent attributes.

The advancement in non-linear learning algorithms and computation power of computers has redirected the missing data imputation approach to machine learning algorithms. The missing data imputation by KNN, an unsupervised learning algorithm, was presented in (Troyanskaya, et al., 2001). It can handle missing instances in multiple attributes without a need for the creation of a separate predictive model for each attribute. However, it suffers from the curse of dimensionality and could be computationally expensive as it searches for similar instances in the entire dataset. Moreover, its hyper-parameter, the number of required clusters, could impact the result. For example, a small number of clusters can provide biased results due to the overemphasis of dominant instances. The imputation of missing values by missForest, a technique based on a random forest algorithm, was proposed by (Stekhoven & Bühlmann, 2012). The researchers chose random forest because it can perform very well under barren conditions such as high dimensions, complex interactions and non-linear data structures. It averages the multiple imputed unpruned classifications or regression trees and estimates the imputation error by built-in out-of-bag error estimates of random forest. A comparative study demonstrated that missForest has a less biased estimate and a narrower confidence interval compared to MICE (Shah, Bartlett, Carpenter, Nicholas, & Hemingway, 2014). A study compared the performance of mean imputation, linear regression, case-wise deletion with ANN, and SVM (Richman, Trafalis, & Adrianto, 2009). This study concludes that case-wise deletion and mean imputation had the largest errors, the performance of linear regression was slightly better than case-wise deletion and mean imputation, and ANN and SVM were the best. The ANN approach was adopted to impute the missing values in attention-deficit hyperactivity disorder data (Cheng, Tseng, Chang, Chang, & Gau, 2020). The results indicated that ANN has higher accuracy and robustness than interpolate imputation, mean imputation, and multiple imputation techniques. The high-performance data imputation techniques based on a machine learning algorithm are black-box in nature. Therefore, the reasoning behind the predicted missing value cannot be explained. The black-box models can be explained by model-agnostic and model-specific techniques (Adadi & Berrada, 2018).

The categorical (qualitative) attributes are transformed into numerical features by encoding techniques, such as one-hot, label, hash, and target encoding. Classical encoding methods assume that categorical attributes are complete; there are no missing values. Therefore, missing values in categorical attributes are imputed before transformation to a required numerical format (discrete labels $\{1,2,3,\ldots\}$ or a one-hot encode $\{0,1\}$). The missForest technique can impute and encode missing categorical attributes into discrete labels. It can be retransformed to a one-hot encode and belief-distribution after imputation if the data is used as an input in ANN and belief-rule-based models (BRB, a type of expert system) (Yang, Liu, Wang, Sii, & Wang, 2006), respectively. The one-hot encode is the most widely used scheme for the numerical representation of categorical attributes (Alkharusi, 2012). Each category of a categorical attribute represents an independent feature, which results in orthogonal vector space equidistance from each other (Cerda, Varoquaux, & Kégl, 2018). A categorical attribute with cardinality $d$ is transformed into $d-$ dimensional vector. The implementation of a one-hot encode in supervised ML methods is easy. It allows learning of each category as a separate parameter in a separate dimension. However, this method is cumbersome when categorical attributes have high cardinality (a large number of categories). It produces a sparse input matrix of numerical data and results in expansion to feature space. The expanded features are problematic for the model training due to long processing time, large space and memory. Additionally, a small or unique number of categories affect performance due to overfitting.

Label encoding is another classical categorical data-encoding technique. It assigns a positive integer to each category of a categorical attribute (Von Eye & Clogg, 1996). Unlike, one-hot encode, it transforms a 1-D vector of a categorical attribute to a 1-D numerical vector. It does not have a dimensionality issue; however, it induces misunderstanding in the learning process of an ML algorithm, especially in the neural network. It assigns higher weight to a larger number and less weight to a small number which skews the model and leads to inaccurate results. This method suits a decision tree and tree-based ensemble methods to find optimal split levels.

Hash encoding is an alternative method to encode categorical data specifically with high cardinality. The basic idea is to assign an integer value to categories and then map integers to another number by a hash function (Weinberger, Dasgupta, Langford, Smola, & Attenberg, 2009). Hashing allows the reduction of dimensionality by a hash function. However, dimensionality reduction causes hash collisions which result in potential loss of information and creation of uninterpretable features. Target encoding methods have received a lot of attention in data science competitions. They map each category in a categorical attribute by the corresponding statistics of its output attribute. The categories are replaced by observed frequency and a mean and estimated probability. However, encoding categories simply by these statistics cause data leakage, which leads to the overfitting of the training set and the poor generation of ML models, primarily due to the presence of a tiny sample of some of the categories. The blend of posterior probability and prior probability as a baseline probability was proposed to mitigate the effect of uncertainty due to the presence of less frequent categories (Micci-Barreca, 2001).

### 2.2 Review on Evidential Reasoning

The evidential reasoning (ER) approach (Yang & Xu, 2002) is based on Dempster–Shafer's theory of evidence (Dempster A. , 2008). The ER approach can deal with the uncertainties such as ambiguities, fuzziness and ignorance in data-driven decision-making. It can provide a decision by combining multiple pieces of independent evidence. A set of continuous numbers representing high-density peak points, an upper and a lower bound in continuous likelihood density function, and a set of categories are set of pieces of evidence in a numerical and a categorical attribute, respectively. The ER approach only considers the weights of evidence. It was extended to ER rule to include the concept of weight and reliability of evidence (Yang & Xu, 2013). The ER rule can combine multiple pieces of independent and highly conflicting evidence with different weight and reliability. It has rigorous and rational probabilistic reasoning process compared to other evidence conjunctive combination rules such as Yager's rule (Yager, 1987), Smets' rule (Smets & Kennes, 1994), Dubois and Prade's rule (Dubois & Prade, 1988), Dempster's rule (Dempster A. , 2008), and proportional conflict redistribution rule (PCR) (Smarandache, Dezert, & Tacnet, 2010).

The concept of reliability of a piece of evidence is used to quantify the quality of information. For decision-making, it is defined as the ability of a piece of evidence from an information source to point correctly to a decision (Smarandache, Dezert, & Tacnet, 2010). Weight of evidence refers to the importance of the evidence which could be subjective in nature to incorporate the decision maker's preferences over a set of the pieces of evidence. The reliability and weight of evidence have different specificity; therefore, they are handled differently in ER rule. It is proven that ER approach and Dempster's rule are a special case of the ER rule, when the reliability of evidence is equal to its weight and when each piece of evidence is fully reliable, respectively (Yang & Xu, 2013). The inferred outcome by ER rule is profiled over subsets in the power set of the frame of discernment. The frame of discernment is a set of mutually exclusive and collectively exhaustive decisions. The inference process implemented in the ER rule has been extended using hybrid models to cover various methods, including the belief rule-based system inference using the ER rule (RIMER) (Almaghrabi, Xu, & Yang, 2019) and a group-oriented paper recommendation method based on probabilistic matrix factorization and evidential reasoning (GPMF_ER) (Wang, Zhang, Wang, Chu, & Shao, 2021). Moreover, the ER rule has been implemented for inference from imperfect data in various applications, such as safety assessments for complex systems (Tang S. W., Zhou, Hu, Zhao, & Cao, 2020), road safety evaluation (Ganji, Abbas Rassafi, & Jamshidi Bandari, 2020), and environmental investment prediction (Yang, et al., 2021). However, it works under the assumption that attributes are independent of each other. The ER rule was extended to maximum likelihood evidential reasoning (MAKER) rule to consider the interrelation between evidence in the input attributes and between the input and output attributes (Yang & Xu, 2017). The interrelation among pair of evidence is measured by marginal and joint probability. The notion of weight and reliability of evidence is identical in both ER and MAKER. However, in MAKER the weight and reliability of each evidence pointing to each subset of propositions in a power set of a frame of discernment. The ER rule considers overall weight and reliability of the evidence; they are considered distinctly for each proposition in a power set of a frame of the discernment. The joint probability mass of joint evidence is obtained from maximum likelihood evidence reasoning (MAKER). The MAKER framework was proposed for inference from ambiguous categorical data (Yang & Xu, 2017) (Liu, Sachan, Yang, & Xu, 2019). It was utilized for the transformation of input data and fusion of attributes for an explainable loan underwriting system based on belief-rule-based (Sachan S. , Yang, Xu, Benavides, & Li, 2020). The fusion of attributes by MAKER rule reduces the number of rules in rule-based systems.

In this paper, MAKER rule is applied for the numerical transformation of an individual incomplete and ambiguous categorical attribute. It is applied to combine two or more categorical attribute to reduce the dimensionality of transformed numerical features. It is especially beneficial for rule-based methods where the number of rules increases exponentially due to a large number of evidence in the attributes. The numerically transformed features can be used as input data in different types of AI/ML models such as tree-based, black box models like ANN, and rule-based models such as BRB.

## 3. Contribution

The development of an automated decision-making system is challenging, and it heavily relies on the credibility of the data. The data used in these systems are extraction from multiple sources. The informational uncertainty in categorical attributes is much higher than numerical attributes (Qin, Xia, & Prabhakar, 2011). The first step to engender human trust in an automated decision-making system demands consideration of uncertainty due to missing values in data and a firm understanding of input data pre-processed by data imputation techniques. The uncertainty challenge by categorical data has not been addressed comprehensively in previous research. This paper attempts to answer the following questions:

Question 1: What are the possible sources of uncertainty in categorical attributes?

Question 2: How does an inadequate understanding of predicted imputed values in categorical attributes affect pre-modelling and post-modelling explainability?

Question 3: How can a missing categorical data imputation technique examine ignorance due to the absence of class labels in an output attribute?

Question 4: How can a missing categorical data imputation technique incorporate subjective human expert opinion?

In this paper, question 1 and question 2 are acknowledged in Section 4, and they are discussed through an application on real data in Section 7.4. Section 4 presents the informational uncertainty, unforeseeable uncertainty in the decision task environment, and the uncertainty due to lack of pre-modelling explainability in categorical data. Missing values in input attributes and output attributes introduce uncertainty in the predicted decisions. Section 5 of this paper presents the maximum likelihood evidential reasoning (MAKER) rule to pre-process the incomplete and ambiguous categorical data. This methodology addresses question 3. A brief review on evidential reasoning is presented in Section 2.2. The MAKER rule presented in Section 5.1 is applied to impute and transform incomplete and ambiguous pieces of evidence in an individual categorical attribute into interpretable, numerical features. The conjunctive MAKER rule presented in Section 5.2 is applied to impute missing values, transform categorical attributes to interpretable, numerical features, and fuse multiple categorical attributes. It can combine multiple pieces of evidence in two or more attributes for dimensionality reduction of numerically transformed features for AI/ML models developed for highly regulated domains. The MAKER for individual attributes and conjunctive MAKER to combine multiple categorical attributes is called I-MAKER and C-MAKER, respectively. MAKER rule methodology can provide reasoning behind predicted missing values, which impart a clear understanding of the input data. This technique is based on the general framework of Dempster Shafter's (DS) theory to consider uncertainty and ignorance caused by incomplete and ambiguous data. The MAKER rule covers the uncertainty in data by considering local ignorance when each evidence in a categorical attribute points to two or more decisions and global ignorance when each evidence in a categorical attribute does not point to any decision; the evidence state is entirely unknown. Due to the absence of class labels, local and global ignorance is considered by profiling the outcome over the power set of the frame of discernment. The uncertainty caused by the importance of the evidence and the sufficiency of the evidence is incorporated in MAKER through the evidence's weight and reliability, respectively. Both I-MAKER and C-MAKER addresses question 4 by demonstrating their ability to assign expert judgment as a subjective weight to each piece of evidence (category) in a categorical attribute. If the subjective weight from experts is not available, then they are trained by the data. The reliability of the evidence is another important parameter in this methodology.

This paper demonstrates the significance of trust in transformed data to obtain explainable and trustworthy decisions. The applicability of the proposed technique is demonstrated in two real datasets on early asthma symptoms and mortgage loans. The performance of three AI models based on ANN (deep learning model), decision tree (tree-based model), and BRB (rule-based model) is compared against a different set of data obtained by the treatment and transformation by I-MAKER and C-MAKER with four other commonly used data imputation techniques. A numerical example in Section 6 demonstrates the step-by-step method to combine two attributes in early asthma symptoms data. Section 7 presents the results and discussion of two case studies on healthcare (based on early asthma symptoms) and finance (based on mortgage loans). The paper concludes in Section 8 presents the advantages of MAKER, the future direction of research, and potential application areas. There are three Appendix. Appendix A shows the algorithmic steps for MAKER for an individual categorical attribute (I-MAKER) and multiple categorical attributes (C-MAKER). Appendix A has additional results from the numerical example presented in Section 6. Appendix C has hyper-parameters of deep-learning and tree-based model and structure of the rule-based system.

## 4. Uncertainty in Decision-Making by Categorical Data

### 4.1 Informational uncertainty

The informational uncertainty arises due to diversity in information (Iselin, 1989), lack of information due to insufficient instances, information loss due to missing values in the data, information representability, and the information source (Wu & Shang, 2020). A categorical attribute can be binary, ordinal, interval or nominal. Lack of information due to insufficient instances, low sample categories in categorical attributes, missing categories in input attributes (incomplete) and missing categories in an output attribute (ambiguity) are the source of informational uncertainty in categorical data. The missing data imputation techniques are based on complete case analysis techniques. The incomplete cases may contain a large amount of information, especially when the number of incomplete cases is relatively larger in comparison to the fully observed cases (Hughes, Heron, Sterne, & Tilling, 2019). The output attribute points specifically to an expected decision for each instance (also called a class label) from a set of possible decisions. A supervised learning system learns from labelled decisions. The unavailability of class labels in an output attribute in a dataset represents the ignorance of pieces of evidence in each attribute to point specifically to a decision. The data imputation techniques fail to consider the ignorance introduced by an output attribute (Baneshi & Talei, 2011) (Baneshi & Talei, 2012). Additionally, in a natural setting, it is presumed that the categories in a categorical attribute are mutually exclusive; that is, there is no overlap between the sets of categories. Poor-quality data could have the problem of overlapping categories. It results in high cardinality and multicollinearity in numerically transformed data. Moreover, categories in an attribute are required to be exhaustive to cover an entire set of possibilities (Micci-Barreca, 2001).

### 4.2 Unforeseeable uncertainty in a decision task environment

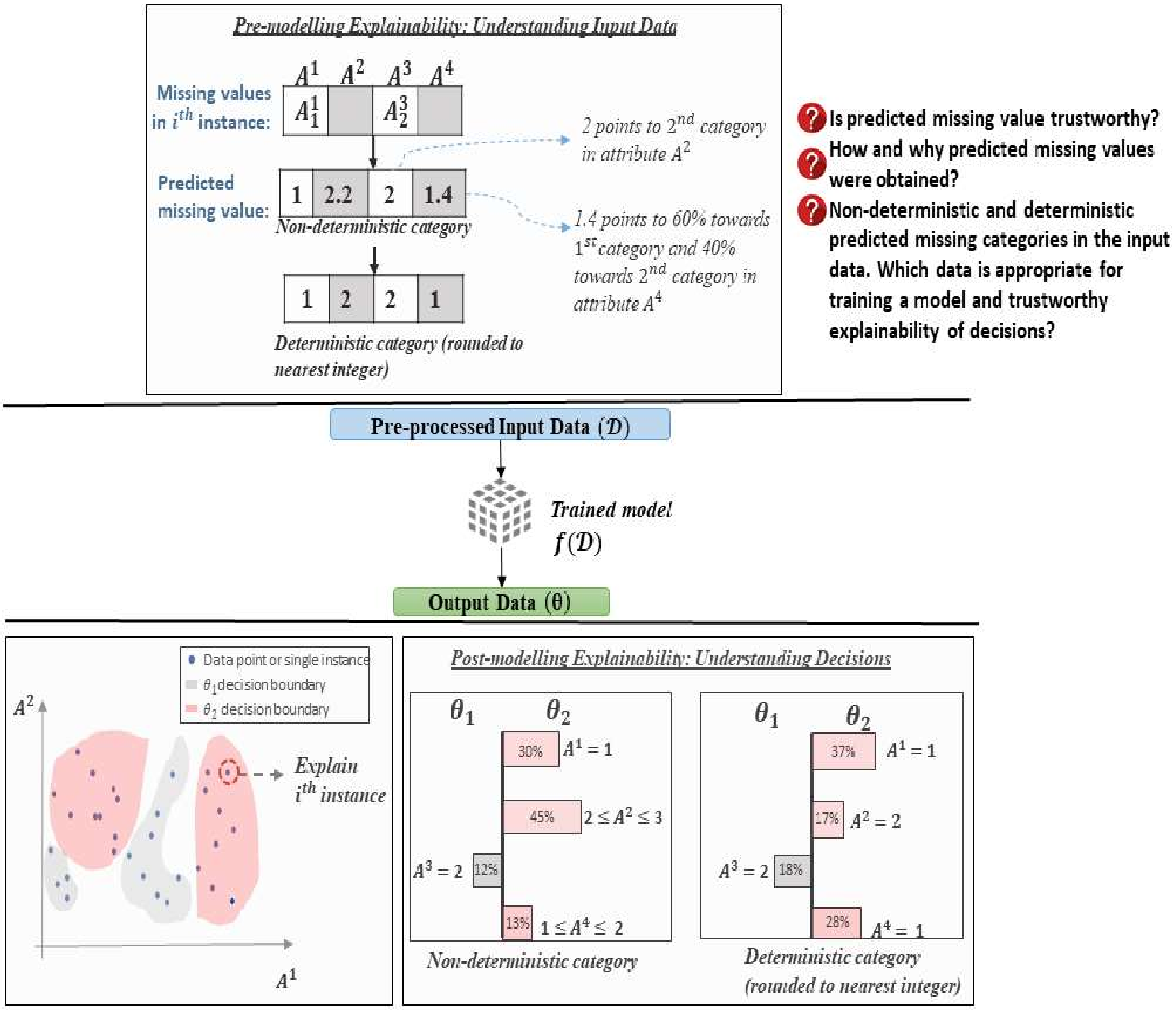
The decision-makers currently face the problem of estimating future uncertainties in data used to train AI-enabled systems. The decision given by an AI system subsists in a specific environment. In the real world, the decision task environment is volatile, which implies that it is continuously changing (Bourgeois III, 1980) (Wu & Shang, 2020). For instance, changes in social trends in marketing data (Ducange, Pecori, & Mezzina, 2018), climate-related changes in disease (Redshaw, Stahl-Timmins, & Fleming, 2013), and policy-related changes in loan lending (Sachan S. , Yang, Xu, Benavides, & Li, 2020). Many attributes in these datasets are categorical to represent sets of evidence. In some applications, numerical attributes are mapped into a certain set of categories (Sachan S. , Yang, Xu, Benavides,

& Li, 2020) (Almaghrabi, Xu, & Yang, 2020). For example, in credit data, the number of defaults or the income is grouped into certain categories to represent a policy or rule within an organisation.

Some of the categories in an attribute do not exist in the collected data and would arrive at some point in the future after the deployment of a decision-making system. The system will fail if the newly arrived category does not exist in a numerically transformed feature space of the categorical attribute. For example, the set categories for attribute $A^1$ is $\{A_1^1, A_2^1, A_3^1\}$. It can be transformed into a 3-D vector $\{[1,0,0], [0,1,0], [0,0,1]\}$ by a one-hot encode. The AI/ML algorithm trained on three features for attribute $A^1$ expects the arrival of one of the three categories in the future. The appearance of a new category would be unknown for the trained model. It can be accommodated by superseding a 3-D vector by one-hot encode with a 4-D vector in numerical transformed data. The algorithm parameters should be retrained with the new data with an extra feature. The problem of the arrival of a new category is not limited to numerical representation by one-hot encoding. It exists in all numerical transformation techniques. Some of the expected unforeseeable uncertainties in categorical attributes could be due to fundamental mismatch between the values predicted for missing data by complete case analysis and a predicted decision by the system in the light of new categories (new evidence), increase number of low sample categories (or unique categories) in the future, and a decrease in high sample categories in the future. Hence, soliciting advice from domain experts in each step for the development of an AI-enabled decision-making system would provide essential knowledge in estimating the future uncertainties for better adaptation in an evolving decision task environment (Bogosian, 2017).

### 4.3 Uncertainty due to lack of pre-modelling explainability

The automated decision-making systems enabled by an AI/ML algorithm learn hidden patterns in the features within a dataset, which are obscure for humans. The potential users and developers must understand the input data utilised for automating the decisions. Understanding input data includes the ability to apprehend missing imputed values predicted by imputation techniques. Some commonly used data imputation techniques are missForest, MICE, EM, and KNN. Among all the techniques, missForest based on random forest algorithm was found most efficient (Stekhoven & Bühlmann, 2012) (Waljee A. , et al., 2013). The deep learning approach based on the ANN algorithm was implemented for missing data imputation (Cheng, Tseng, Chang, Chang, & Gau, 2020) (Richman, Trafalis, & Adrianto, 2009). The results indicated that deep learning provides higher accuracy than tradition statistical imputation methods. Both ANN and random forest are black-box in nature. The predicted missing values by these models cannot be explained directly unless a model agnostic method to leverage an inherently interpretable surrogate model or a model-specific method is used to explain the outcome (Kelly L. , et al., 2020). All missing data imputation techniques predict missing values by analysing complete cases and avoiding incomplete cases. The predicted missing values by these techniques could be probabilistic; for example, 0.60 for category $A_1^1$ and 0.40 for category $A_2^1$ in attribute $A^1$. It could be a continuous value; for example, an imputed missing value 1.4 points 60% towards category $A_1^1 = 1$ and 40% towards category $A_2^1 = 2$ in an attribute $A^1$. The predicted category is naturally converted into a numerical value; however, it may not point to a concrete category. It could point to a specific category if the predicted value is rounded to a nearest integer, which points to a category in a categorical attribute. Therefore, the inexplicability of missing imputed data and the inability to estimate a definite category are two main issues of high-performance machine learning-based missing data imputation techniques. Human understanding for a decision (output) given by the system is dependent on the initial understanding of the input data because a decision by AI/ML model is explained in terms of contribution of the features in the data. The inability to explain how a missing value was predicted in an attribute and to which category the predicted value explicitly points to would provide an unreliable explanation of the decision. The pre-modelling explainability of the data is important for trustworthy post modelling explainability of the decisions. It is dangerous to allow AI to take charge until an adequate understanding of the contextual structure decision process is accomplished by human experts. A recent article in Harvard Business Review has stressed the importance of understanding input data (Agrawal, Gans, & Goldfarb, 2020). Figure 2 illustrates how a lack of understanding of the input data affects the explainability of a decision in the post-modelling stage.



Pre-modelling Explainability: Understanding Input Data
$A^1$ $A^2$ $A^3$ $A^4$
Missing values in $i^{th}$ instance:
$A_1^1$
$A_2^3$
2 points to $2^{nd}$ category in attribute $A^2$
Predicted missing value:
1 2.2 2 1.4
Non-deterministic category
1.4 points to 60% towards $1^{st}$ category and 40% towards $2^{nd}$ category in attribute $A^4$
1 2 2 1
Deterministic category (rounded to nearest integer)
Is predicted missing value trustworthy?
How and why predicted missing values were obtained?
Non-deterministic and deterministic predicted missing categories in the input data. Which data is appropriate for training a model and trustworthy explainability of decisions?
Pre-processed Input Data ($\mathcal{D}$)
Trained model $f(\mathcal{D})$
Output Data ($\theta$)
Data point or single instance
$\theta_1$ decision boundary
$\theta_2$ decision boundary
$A^2$
$A^1$
Explain $i^{th}$ instance
Post-modelling Explainability: Understanding Decisions
$\theta_1$ $\theta_2$
30% $A^1 = 1$
45% $2 \leq A^2 \leq 3$
$A^3 = 2$ 12%
13% $1 \leq A^4 \leq 2$
Non-deterministic category
$\theta_1$ $\theta_2$
37% $A^1 = 1$
17% $A^2 = 2$
$A^3 = 2$ 18%
28% $A^4 = 1$
Deterministic category (rounded to nearest integer)

Figure 2. Uncertainty due to lack of pre-modelling explainability in missing categorical data

## 5. Methodology

### 5.1 I-MAKER: MAKER Rule to Pre-Process Individual Categorical Attribute

Let a dataset be denoted by $\mathcal{D}$. It can have both numerical and categorical attributes. For simplicity, let's assume that $\mathcal{D}$ has $q \in \{1, \ldots, Q\}$ number of categorical attributes and $v \in \{1, \ldots, V_q\}$ number of categories in each attribute. An output attribute is denoted by $\theta$. The frame of discernment is given by $\Theta = \{\theta_1, \ldots, \theta_z, \ldots, \theta_Z,\ z \in \{1, \ldots, Z\}\}$. It is a mutually exclusive and collectively exhaustive set containing all possible class labels in output attributes. A class label represents an expected decision from a model. Therefore, a set of class labels in an output attribute can also be called a set of decisions. A dataset $\mathcal{D}$ is shown in Table 1. For simplicity, it is assumed that dataset ($\mathcal{D}$) contains only categorical attributes.

Table 1: Data ($\boldsymbol{\mathcal{D}}$)

| $data$ $point$ | $\boldsymbol{A^1}$ | … | $\boldsymbol{A^q}$ | … | $\boldsymbol{A^Q}$ | $\boldsymbol{\theta}$ |
|---|---|---|---|---|---|---|
| 1 | $A_v^1$ | | $A_v^q$ | | $unknown$ | $\theta_z$ |
| ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ |
| $i$ | $unknown$ | | $A_v^q$ | | $A_v^Q$ | $unknown$ |
| ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ |
| $I$ | $A_v^1$ | | $A_v^q$ | | $A_v^Q$ | $\theta_z$ |

The power set of the frame of discernment $(\Theta)$ consists of $2^Z$ subsets of $\Theta$. It is denoted by $2^\Theta$ or $P(\Theta)$. It can be written as:

$$2^\Theta = P(\Theta) = \{\emptyset, \{\theta_1\}, \dots, \{\theta_Z\}, \{\theta_1, \theta_2\}, \dots, \{\theta_1, \theta_Z\} \dots, \{\theta_1, \dots, \theta_{Z-1}\}, \Theta\} \tag{1}$$

The classes of the output attribute in a classification algorithm are deterministic as the number of the classes in a model is known in advance. The missing class labels can be measured as $unknown = \Theta = \{\theta_1, \dots, \theta_z, \dots, \theta_Z,\ z \in \{1, \dots, Z\}\}$, which shows that the missing label could belong to any class in the set $\Theta$. A set of class labels represents a set of possible decisions. This method maps data into numerical features representing singleton (set with exactly one element) subsets $\{\theta_1\}, \dots, \{\theta_z\}$ in power set. The remaining subsets of power set are empty $(\emptyset)$ because usually, an $i^{th}$ instance (a row in dataset $\mathcal{D}$) does not point to two or more than two classes in an output attribute. Therefore, the power set $P(\Theta)$ for MAKER can be rewritten as follows:

$$P(\Theta) = \{\{\theta_1\}, \dots \{\theta_z\}, \dots, \{\theta_Z\}, \Theta\} \tag{2}$$

Similarly, the missing values in input attributes can be marked as *'Unknown'* where *'Unknown'* is a mutually exclusive and collectively exhaustive set of all deterministic categories (i.e. existing known categories). If categories are not exhaustive (in other words non-deterministic), then the missing antecedent attributes can still be marked as *'Unknown'*. However, in this case, it will be a mutually exclusive and collectively exhaustive set of existing known and future unknown categories. For example, a categorical attribute $A^1$ has three known categories and some missing values. The set of possible categories in $A^1$ could be $\mathbb{A}^1 = \{A_1^1, A_2^1, A_3^1, Unknown\}$. In this set $A_1^1, A_2^1$, and $A_3^1$ are known categories. The missing values is represented by $Unknown = \{A_1^1, A_2^1, A_3^1\}$ if set of known categories are deterministic and $Unknown = \{A_1^1, A_2^1, A_3^1, future\ values\}$ if known categories are non-deterministic. In Table 1, the missing values in input attribute $(A^q)$ and an output attribute $(\theta)$ are marked as *'Unknown'*.

- *Definition of evidence:* Evidence is denoted by $e$ in general. The $v^{th}$ evidence from the $q^{th}$ attribute is denoted by $e_{v,q}$, representing that the $q^{th}$ attribute takes the value of its $v^{th}$ category.

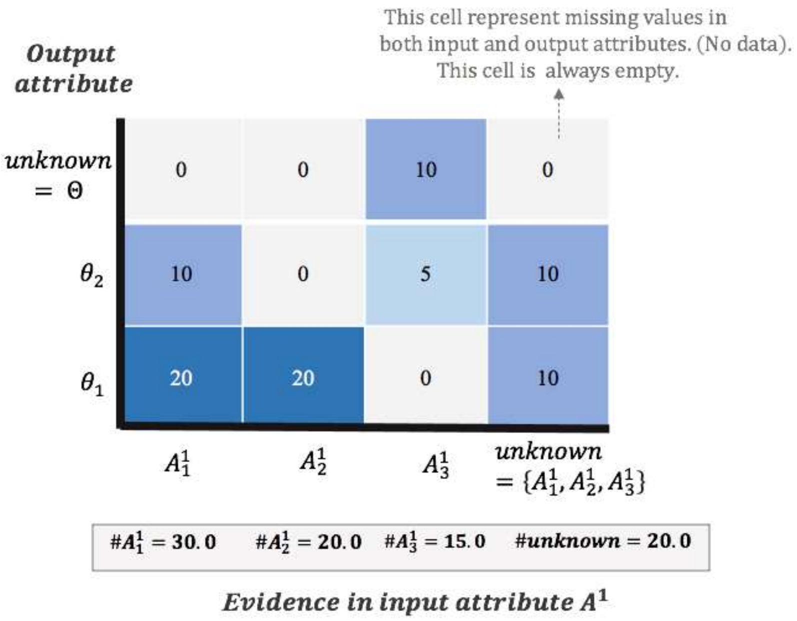
Imprecision in classification occurs due to uncertainty caused by the fact that all samples for a piece of evidence do not point to a specific class (or decision). For example, Figure 3 there are 30 samples for category $A_1^1$ where, twenty samples are pointing to class $\theta_1$, ten samples are pointing to class $\theta_2$, and there are zero samples for unknown values, $\Theta = \emptyset$ . The uncertainty is measured by number of samples which support a class. The extent of support is called belief, obtained from estimating probability mass for each evidence for a given class. Evidence $e_{v,q}$ can be represented as belief-distribution as follows (Yang & Xu, 2017):

$$e_{v,q} = \left\{\left(e_{\theta,v,q}, \widehat{m}_{\theta,v,q}\right), \forall \theta \in P(\Theta)\right\} \tag{3}$$

$$\textstyle\sum_{\theta \in P(\Theta)} \widehat{m}_{\theta,v,q} = 1 \tag{4}$$

In Expression (3) and (4), $e_{\theta,v,q}$ represents a $v^{th}$ category of an $q^{th}$ attribute points to a class $\theta$ and $\widehat{m}_{\theta,v,q}$ is the normalised probability mass for the $v^{th}$ category in an $q^{th}$ attribute points to a class $\theta$. The probability mass of category for a given class is obtained by considering the weight, reliability, and basic probability of evidence for a given class.  Basic probability is obtained by first generating a contingency table. A contingency table is a frequency distribution table, as example is shown in Figure 3. A division operation in each cell of the contingency table by its sum of rows is performed to obtain the likelihood of each evidence in an attribute for a given class. The likelihood of observing the $v^{th}$ category of the $q^{th}$ categorical attribute for a given class $\theta$ is denoted by $L_{\theta,v,q}$. Basic probability $p_{\theta,v,q}$ is calculated by normalising likelihood for all classes $\theta \in P(\Theta)$ as follows (Yang & Xu, 2017):

$$p_{\theta,v,q} = \frac{L_{\theta,v,q}}{\sum_{\theta \in P(\Theta)} L_{\theta,v,q}}, \forall \theta \in P(\Theta) \tag{5}$$



This cell represent missing values in both input and output attributes. (No data). This cell is always empty.

| Output attribute | $A_1^1$ | $A_2^1$ | $A_3^1$ | $unknown = \{A_1^1, A_2^1, A_3^1\}$ |
|---|---|---|---|---|
| $unknown = \Theta$ | 0 | 0 | 10 | 0 |
| $\theta_2$ | 10 | 0 | 5 | 10 |
| $\theta_1$ | 20 | 20 | 0 | 10 |

$\#A_1^1 = 30.0$ $\#A_2^1 = 20.0$ $\#A_3^1 = 15.0$ $\#unknown = 20.0$

Figure 3. An example of a contingency table

The weighted basic probability is referred to as probability mass, denoted by $m_{\theta,v,q}$. In ER rule, the non-normalised probability mass is discounted by the weight of the evidence ($w_{v,q}$) (Yang & Xu, 2013). In MAKER rule, the non-normalised is discounted by the weight of the evidence for a given class $\theta$ ($w_{\theta,v,q}$) (Yang & Xu, 2017). It is the measure of a degree of support for evidence $e_{v,q}$ for a class $\theta$, given as follows:

$$m_{\theta,v,q} = w_{\theta,v,q}\, p_{\theta,v,q} \tag{6}$$

- *Definition of weight of the evidence:* The weight of evidence ($w_{\theta,v,q}$) is the relative importance of evidence in comparison with other evidence. It could be a subjective judgment of experts. If the expert's opinion is not available, then it could be treated as a parameter which could be trained by data-driven optimisation.
- *Definition of reliability of the evidence:* The reliability of evidence ($r_{\theta,v,q}$) is the ability of evidence to point correctly to a class. It is the classification ability of a category in an attribute. A piece of evidence

$(e_{v,q})$ from an attribute $A^q$ is most reliable when it has the most samples for a particular class $(\theta)$. The reliability of a piece of evidence is 1 if all (100%) samples point towards a particular class $(\theta)$. The reliability of a category in an attribute can be calculated by (Xu, Zheng, Yang, Xu, & Chen, 2017):

$$r_{\theta,v,q} = \frac{\#e_{\theta,v,q}}{max_{\theta \in P(\Theta)} \#e_{\theta,v,q}}, \quad \forall \theta \in P(\Theta) \tag{7}$$

where $\#e_{\theta,v,q}$ is the number of samples available for evidence $e_{v,q}$ for class $\theta$. An evidence is most reliable for a class $\theta$ if, $r_{\theta,v,q} = 1$. The reliability of evidence pointing to other classes is relative. Figure 3 demonstrates the number of samples in each category $\{A_1^1, A_2^1, A_3^1, Unknown\}$ of input attribute $A^1$ for all class $\theta \in P(\Theta)$ where, $P(\Theta) = \{\{\theta_1\}, \{\theta_2\}, \Theta\}$. A dark colour in a column represents the class for which a piece of given evidence is most reliable, and a dark colour in a row represents the most reliable evidence for a given class. The probability mass $(m_{\theta,v,q})$ is normalised by weight and reliability as follows:

$$\widehat{m}_{\theta,v,q} = \begin{cases} 0 & \theta = \emptyset \\ c_{\theta,v,q} m_{\theta,v,q} & \theta \subseteq \Theta, \theta \neq \emptyset \\ c_{\theta,v,q}\left(1 - r_{\theta,v,q}\right) & \theta = P(\Theta) \end{cases} \tag{8}$$

In Expression (8), the probability mass is normalised by factor $c_{\theta,v,q} = 1/(1 + w_{\theta,v,q} - r_{\theta,v,q})$ and $\sum_{\theta \in P(\Theta)} \widehat{m}_{\theta,v,q} + \widehat{m}_{P(\Theta),v,q} = 1$. The unreliability of evidence is residual support $\left(1 - r_{\theta,v,q}\right)$. It is earmarked to the powerset consisting of subsets of the frame of discernment, instead of assigning it to the entire frame of discernment. If $w_{\theta,v,q} = w_{v,q}$ and $r_{\theta,v,q} = r_{v,q}$ then it reduces to ER rule (Yang & Xu, 2013).

The weight of the evidence for a given class is trained if a subjective judgment by domain experts is not available. Before training, the initial weight can be set equal to its reliability $(w_{\theta,v,q} = r_{\theta,v,q})$. The training can be avoided for fast pre-processing of incomplete and ambiguous categorical data if it is assumed that evidence with the highest reliability could have relatively high importance compared to other evidence. The objective function for training of weight for each piece of evidence for a class $(w_{\theta,v,q})$ is:

$$\begin{gathered} Minimize{:}f\left(w_{\theta,v,q}\right) = \frac{1}{2I} \sum_{i=1}^{I} \sum_{\theta \in P(\Theta)} \left(m^o - \widehat{m}(w_{\theta,v,q})\right)^2 \\ constraints{:}0 \leq w_{\theta,v,q} \leq 1 \end{gathered} \tag{9}$$

where $i \in \{1, \ldots, I\}$ is an instance in a dataset $\mathcal{D}$. The observed probability for an instance $i$ is denoted by $m^o$ and $\widehat{m}(w_{\theta,v,q})$ is the estimated normalised probability mass.

The table for normalised probability mass $(\widehat{m_{\theta,v,q}})$ is used to map a categorical attribute into $Z$ or $Z + 1$ dimensional numerical feature space when an output attribute has complete or missing classes, respectively. Each column of a categorical attribute in a dataset $\mathcal{D}$ is transformed into $Z$ or $Z + 1$ number of columns. Each column of the numerically transformed feature of a categorical attribute represents the subsets in power set $P(\Theta)$ shown in Equation (2). Naturally, if classes in the output attribute are not missing or are complete, then $Unknown = \Theta = \emptyset$. Each row represents the probability mass for a category in a categorical attribute distributed over subsets in power set $P(\Theta)$ (representation of a piece of evidence in the form of distribution by Expression (3) and (4)). Figure 4 demonstrates the mapping of each category to the corresponding estimated probability mass for attribute $A^1$ in a dataset $\mathcal{D}$. The estimated probability mass for each category $\{A_1^1, A_2^1, A_3^1, Unknown\}$ for a given proposition in $P(\Theta)$ is shown in Table A, Figure 4. The data in Table B in Figure 4 represents the categories for each instance for the attribute $A^1$, and these categories are mapped to the corresponding estimated probability mass from Table A. The resulting transformed data does not lose interpretation as each row in Table B in Figure 4 represents the probability mass for a missing or non-missing category. The estimated probability mass can be explained in terms of two parameters, reliability and weight of a category (a piece of evidence) for a given outcome $\theta \in P(\Theta)$.

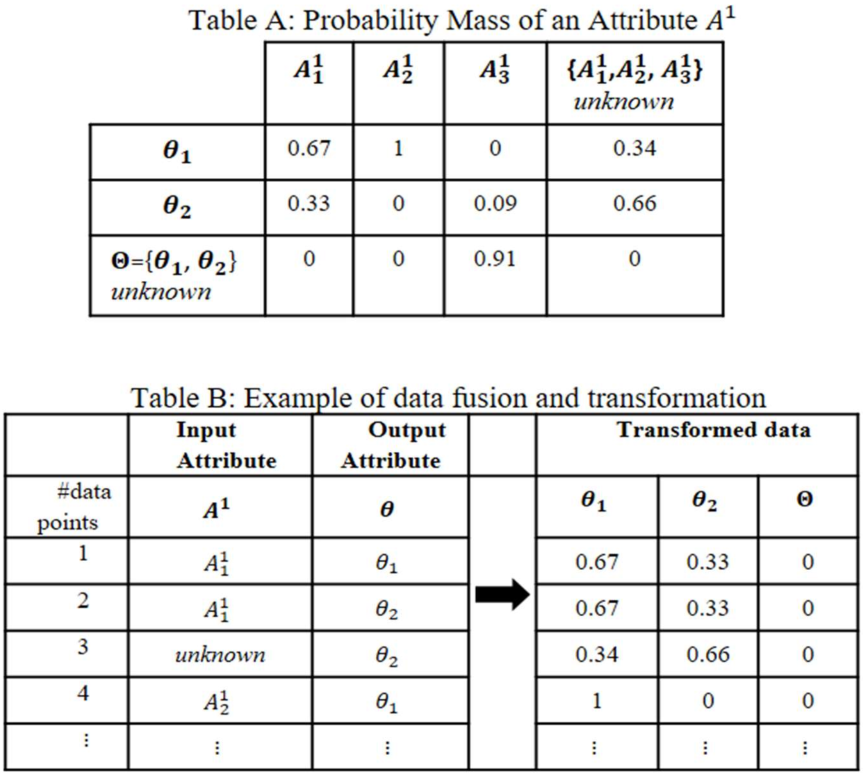


Table A: Probability Mass of an Attribute $A^1$

| | $A_1^1$ | $A_2^1$ | $A_3^1$ | $\{A_1^1, A_2^1, A_3^1\}$<br>*unknown* |
|---|---|---|---|---|
| $\theta_1$ | 0.67 | 1 | 0 | 0.34 |
| $\theta_2$ | 0.33 | 0 | 0.09 | 0.66 |
| $\Theta=\{\theta_1, \theta_2\}$<br>*unknown* | 0 | 0 | 0.91 | 0 |

Table B: Example of data fusion and transformation

| | Input Attribute | Output Attribute | | Transformed data | | |
|---|---|---|---|---|---|---|
| #data points | $A^1$ | $\theta$ | | $\theta_1$ | $\theta_2$ | $\Theta$ |
| 1 | $A_1^1$ | $\theta_1$ | ➡ | 0.67 | 0.33 | 0 |
| 2 | $A_1^1$ | $\theta_2$ | | 0.67 | 0.33 | 0 |
| 3 | *unknown* | $\theta_2$ | | 0.34 | 0.66 | 0 |
| 4 | $A_2^1$ | $\theta_1$ | | 1 | 0 | 0 |
| ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ |

Figure 4. An example of interpretable transformation of attribute $A^1$ by I-MAKER

The MAKER rule is beneficial in pre-processing incomplete and ambiguous data. It can transform categories in an attribute into interpretable numerical features by considering ignorance due to missing values in input and output attributes. It is a type of interpretable machine learning algorithm. The missing value prediction by a machine learning model has a tendency for data leakage (Schelter, et al., 2020). It leads to overfitting of training data and a poor fit of validation data, which confirms poor generalisation of an AI-enabled decision-making system. It can leak information from an output attribute ($\theta$) to numerically transformed categorical attributes ($A^q$). The $k$-fold cross-validation can be used for regularisation to suppress data leakage and overfitting. The less frequent categories get more randomised compared to the most frequent categories. For $k$-fold regularisation of the MAKER pre-processing method, the randomly permuted sample of data is split into two parts, cross-validation (CV) set and validation set. The proportion of data in cross-validation set depends on the size of the data. The CV set is split into $K$ equal parts (also called folds). The probability mass $\widehat{m}_{\theta,v,q}$ is calculated for all $A^q$ ($q \in \{1, \dots, Q\}$) categorical attributes in each $K-1$ folds, excluding one-fold for testing. The average probability mass calculated from $K$ different samples is used to transform the categorical attribute in the test fold. The prior is used as an independent piece of evidence in MAKER. It is used to update the probability mass of missing values in attributes labelled as $unknown = \left\{A_1^q, \dots, A_v^q, \dots, A_{V_q}^q\right\}$. The probability mass of $unknown$ can be adjusted by combining the posterior with the prior. Dirichlet smoothing is widely used to overcome the under-sampling problem (Han, Jiao, & Weissman, 2015). The posterior probability mass of $unknown$ pointing to a class $\theta$ and prior pointing to a class $\theta$ are combined together by Dirichlet smoothing (Micci-Barreca, 2001) (Simonoff, 1995). The posterior probability can be written as $\widehat{m}_{\theta,unknown,q}$ , $\forall\theta \in P(\Theta)$ and the prior can be written as $Prior_{\theta,q}$. The prior for a category for a given class $\theta$ is the proportion of the number for samples for the categories pointing to $\theta$ by the total number samples in a dataset $\mathcal{D}$. The total number of samples in a dataset is equal to the number of instances ($I$). The prior is given as:

$$Prior_{\theta,q} = \frac{\#e_q(\theta)}{I}, \forall\theta \in P(\Theta) \tag{10}$$

where $\#e_q(\theta)$ is all categories in an attribute $q$ pointing to $\theta$, $\forall\theta \in P(\Theta)$. The ability of the posterior to predict the specific class for a missing value is found by its accuracy. The accuracy of $unknown$ to point to a class $\theta$ ($\theta \in P(\Theta)$) is denoted by $\lambda_{\theta,q}$. The prior and posterior can be combined by Dirichlet smoothing by the following equation (Micci-Barreca, 2001):

$$\widehat{m}^{s}{}_{\theta,unknown,q} = \lambda_{\theta,q}\, \widehat{m}_{\theta,unknown,q} + (1 - \lambda_{\theta,q})\, Prior_{\theta,q} \tag{11}$$

A probability mass ($\widehat{m}^{s}{}_{\theta,unknown,q}$) converges towards prior if the accuracy of the posterior is very poor and vice versa. The probability mass of $Unknown$ is adjusted by Equation (11). It is important to note that if categories in a categorical attribute are non-deterministic (non-exhaustive) and future categories are independent of existing categories, then only prior probability must be used for missing values. The steps to pre-process categorical attributes by the MAKER framework for the transformation of an individual categorical attribute is shown in Table 16, Appendix A. The MAKER framework for an individual categorical attribute is denoted as I-MAKER.

## 5.2 C-MAKER: Conjunctive MAKER to Pre-Process and Combine Multiple Categorical Attributes

### 5.2.1 Sufficent Statistics to Combine Attributes

Two or more attributes can be combined when evidence (categories) in attributes are interdependent and enough data is available to generate joint probability mass. The Chi-square test for a contingency table could be used to check the dependency in the joint pieces of evidence. However, some of the joint pieces of evidence may not be available in real-world datasets. The Chi-square test requires at least five samples at each cell in the contingency table (Goodman, 1971) (Fisher, 1992). The Chi-square test is irrelevant for cells with small or no values. The Fisher exact test is from the family of Chi-square tests. Like the Chi-square test, it can examine the association between two different elements in rows and columns of a contingency table. It is based on hypergeometric distribution, and there is no lower limit for the number of samples in each cell of the contingency table (Fleiss, Levin, & Paik, 2013). It is computationally expensive for a large contingency table and is commonly applied to a $2 \times 2$ sized contingency table.

This section presents an alternative approach, based on the interdependence index and the evidence sparsity index, to estimate the potential of combining two or more attributes in a dataset ($\mathcal{D}$). These tests are performed on small ($\geq 2$) to large numbers of attributes. If there are $Q$ number of categorical attributes, then tests can be performed to check the potential of combining $g$ ($g \in \{2\ to\ G\}$) number of categorical attributes into a single attribute. A small set of attributes that satisfy interrelation conditions are fused together. Suppose, $G$ is the maximum number of attributes that are potentially combined into a single attribute ($G < Q$). If $G = Q$, then the MAKER framework could be used for inference, like any other machine learning model. The inference refers to the process of training and prediction. In this research, it is adopted for transformation and fusion of input data for AI/ML algorithms such as rule-based, deep-learning, and tree-based. The potential of combining two attributes ($g = 2$) is tested first $\{A^1, A^2\}, \{A^1, A^3\}, \dots, \{A^q, A^{q+1}\}, \dots,\ \{A^{Q-1}, A^Q\}$, then three ($g = 3$) attributes $\{A^1, A^2, A^3\}, \dots\ , \{A^q, A^{q+1}, A^{q+2}\}, \dots, \{A^{Q-2}, A^{Q-1}, A^Q\}$ as so on. The total number of combinations that can be created from the attributes is given by:

$$\Omega = \sum_{g=2}^{G} \frac{Q!}{g!\,(Q-g)!} \tag{12}$$

In expression (12), $\Omega$ (upper case omega) represents the total number of combinations of the attributes, such that $\omega =$ 1 to $\Omega$. Following are two types of tests – interrelation index and evidence sparse index - to evaluate the suitability of combining two or more categorical attributes by conjunctive MAKER. The conjunctive MAKER can transform and fuse categorical attributes into numerical features. It can impute and transform missing values by analysing entire data which considers the uncertainty induced due to incomplete & ambiguous data. Feasible combinations of categorical attributes have a high interrelation index and a high sparse index. The set of feasible combinations of attributes is represented by $\overline{\overline{\Omega}}$, where the cardinality of set $\overline{\overline{\Omega}}$ is less than $\Omega$ and each combination in this set contains attributes without duplication. For example, suppose a dataset has $Q = 13$ number of categorical attributes and the potential of combining a maximum of $G = 3$ number of categorical attributes is tested. Then, the total number of combinations of all $Q = 13$ attributes in a group containing a minimum of $g = 2$ and a maximum of $G = 3$ attributes would be 78 and 286, respectively. Therefore, the total number of combinations would be $\Omega = 364$. For instance, through the sufficient statistics tests, it is found that there are three most feasible combinations of categorical attributes out of $\Omega = 364$ combinations: $\overline{\overline{\Omega}} = \{\{A^1, A^2\}, \{\{A^4, A^7\}, \{A^3, A^{11}, A^{13}\}\}$. The attributes selected for the fusion by conjunctive MAKER could be suggested by the experts or reflect the prevalent domain practices. If domain knowledge is not available, then a set of attributes can be combined by following the interrelation index and sparse index tests.

***5.2.1.1 Interrelation Index (ψ):*** The interrelation between two pieces of evidence from two different attributes is obtained through their single and joint probability for a given class $\theta$, where $\theta \in P(\Theta)$. Suppose, a set of evidences (or categories) in attribute $A^1$ and $A^2$ is $\mathbb{A}^1 = \{\{A_1^1\}, \{A_2^1\}, unknown\}$ and $\mathbb{A}^2 = \{\{A_1^2\}, \{A_2^2\}, unknown\}$, respectively. The $v^{th}$ evidence in $\mathbb{A}^1$ and the $v'^{th}$ evidence in $\mathbb{A}^2$ are $v \in \{1, \dots, V_1 = 3\}$ and $v' \in \{1, \dots, V_2 = 3\}$, respectively. The missing values $A^1$ and $A^2$ are represented by $unknown = \{A_1^1, A_2^1\}$ and $unknown = \{A_1^2, A_2^2\}$, respectively. Similarly, missing values in an output attribute are represented as $unknown = \{\theta_1, \theta_2\}$, here $\{\theta_1, \theta_2\} \in P(\Theta)$.

The interrelation index between two pieces of evidence is obtained from joint basic probability between two pieces of evidence and singleton probability of each evidence. The basic probability of evidences in individual attributes $A^1$, $A^2$ and both attributes $A^1\ and\ A^2$ pointing to a class $(\theta, \forall\theta \in P(\Theta))$ is denoted by $p_{\theta,v,1}$, $p_{\theta,v',2}$, and $p_{\theta,v1,v'2}$, respectively. The basic probability is the normalised likelihood, which is obtained from the contingency table for pieces of evidence in attribute $A^1$, $A^2$, and both attributes $A^1\ and\ A^2$ for its corresponding class $(\theta)$ from complete cases in dataset $\mathcal{D}$, demonstrated in Figure 4. A complete dataset is denoted by $\mathcal{D}_{CM}$, and $\mathcal{D}_{CM} = \mathcal{D}$ when input attributes in dataset $\mathcal{D}$ have no missing values. The number of joint evidences of multiple attributes is equal to the cartesian product of the number of categories (or single evidences) in all $Q'$ number of categorical attributes.

$$c = \prod_{q=1}^{Q'} V_q \tag{13}$$

The basic probabilities are calculated using Equation (5). The interrelation index between two evidences $e_{v,A^1}$ and $e_{v',A^2}$ pointing to class $\theta_1$ and $\theta_2$, with $\theta_1 \cap \theta_2 = \theta, \forall\theta \in P(\Theta)$, respectively, is given as follows:

$$\psi_{\theta,v1,v'2} = \begin{cases} 0 \quad if\ p_{\theta_1,v,1} = 0\ or\ p_{\theta_2,v',2} = 0 \\ \dfrac{p_{\theta,v1,v_2'}}{p_{\theta_1,v,1}\ p_{\theta_2,v',2}} \qquad otherwise \end{cases} \tag{14a}$$

where, $\psi_{\theta,v1,v_2'}$ is the interrelation index of a joint evidence $e_{v1,v_2'}$ pointing to a class $\theta$, $\forall\theta \in P(\Theta)$. The interrelation index is considered not defined, if individual probabilities are $p_{\theta_1,v,1} = 0$ and $p_{\theta_2,v',2} = 0$. The interrelation index has the following property:

$$\psi_{\theta,v1,v_2'} = \begin{cases} 0, & disjoint \\ 1, & independent \end{cases} \tag{14b}$$

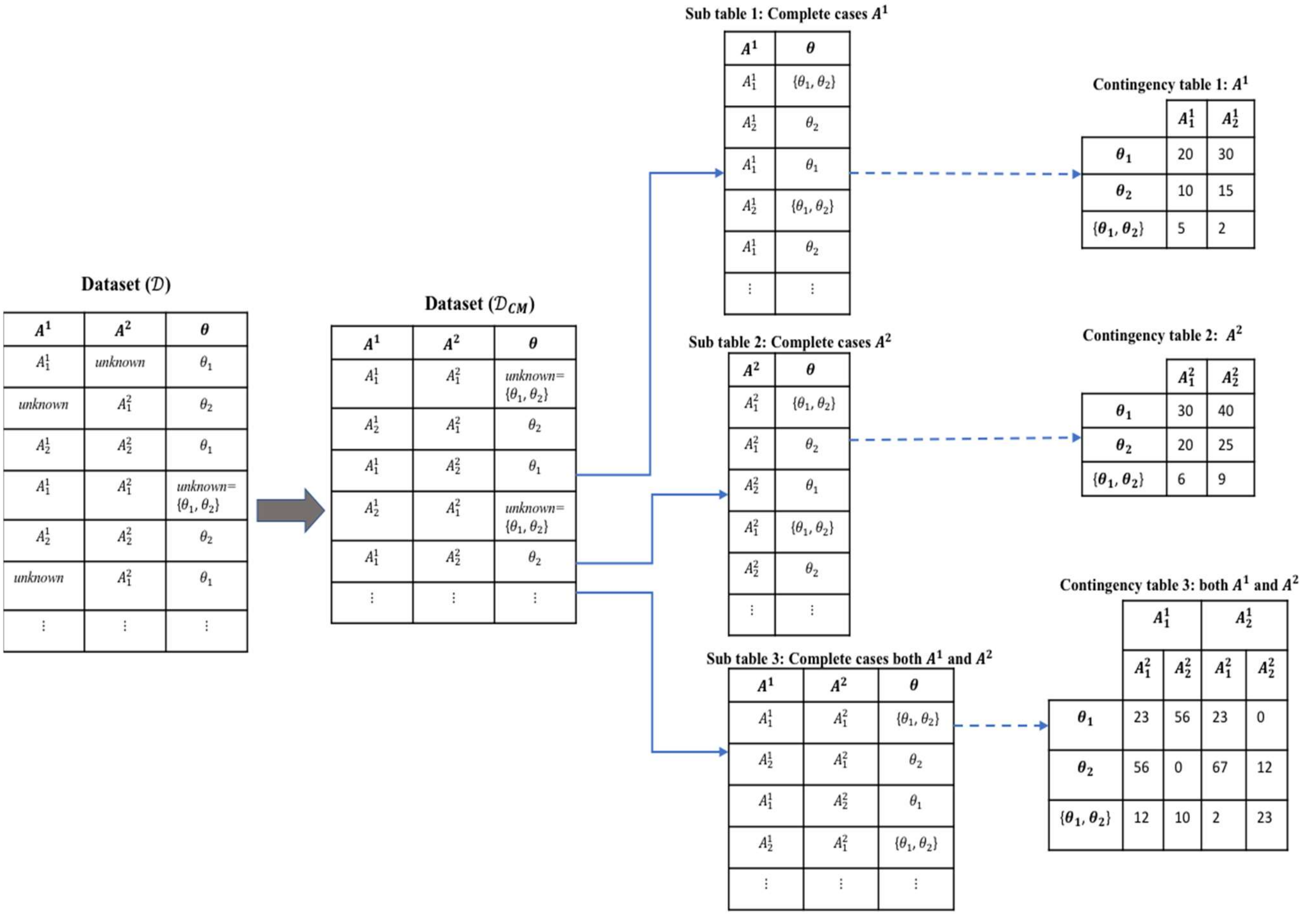
Two evidences are interrelated when $\psi_{\theta,v1,v_2'} \neq 1$. In Equation (14b), the disjoint evidences are always dependent. The disjoint is extreme dependence, where the occurrence of one evidence for a class $\theta$ conveys information that other evidence will not occur in an identical class $\theta$.

The interrelation between two pieces of evidence in two attributes can be represented by the table shown in Figure 6. The cells in an interrelation index table can have three types of values. A cell may not have an interrelation index $\psi = null$ due to an absence of instances for joint pieces of evidence in the contingency table obtained from $\mathcal{D}_{CM}$. For example, there are zero samples for joint pieces of evidence for $A_1^1$ and $A_2^2$ for $\theta_2$ in contingency table 3 in Figure 5. A cell in an interrelation table can have interrelated pieces of evidence when $\psi \neq 1$ and $\psi = 0$. It can have independent pieces of evidence when $\psi = 1$. The number of columns and type of joint evidences in a contingency table between different attributes vary due to the existence of different types of evidence in these attributes. The degree of interrelation between evidences in multiple interrelation tables can be compared by the proportion of interrelated evidences. The proportion of interrelated pieces of evidence in two attributes $\{A^1, A^2\}$ is given by:

$$\psi^{p}{}_{1,2} = \frac{\hat{\psi}_{\theta,v1,v_2'}}{\prod_{q=1}^{2} V_q} \tag{15}$$

where, $\psi^p$ denotes the proportion of interrelated evidences in attributes $\{A^1, A^2\}$, $\hat{\psi}_{\theta,v1,v_2'}$ is the number of evidences having an interrelation index not equal to one and equal to zero ($\psi \neq 1$ $and$ $\psi = 0$), and $\prod_{q=1}^{2} V_q$ is the number of joint evidences (Equation (13)). If all evidences are dependent, then $\hat{\psi}_{\theta,v1,v_2'} = \prod_{q=1}^{2} V_q$ and $\psi^{p}{}_{1,2} = 1$. The empty cell in the contingency table will have the value $\psi_{\theta,v1,v_2'} = null$; a high sparsity in the contingency table would result in a small value of $\psi^{p}{}_{1,2}$.

Each column represents a joint piece of evidence such as contingency table 3 in Figure 5 and the interrelation index table in Figure 6. Two or more pieces of evidence in two or more attributes can be combined by conjunctive MAKER. For example, if there are three attributes, then any two attributes are combined together before combining them with third attribute. It is an iterative process, will be explained in detail in the next section. Furthermore, the highest joint probability of a joint piece of evidence pointing towards a specific class ($\theta \in P(\theta)$) indicates high density towards it, instead of scattered density towards all classes that does not point specifically towards any particular outcome with confidence. Therefore, potentially attributes can be combined when average of all maximum joint probability for joint pieces of evidence towards a class $\left(\theta \in P(\Theta)\right)$ in interrelated evidences ($\hat{\psi}_{\theta,v1,v'2} \neq 1$ or $=$ 0) has a high value. The average of the maximum joint probability of joint pieces of evidence towards a class is denoted by $\mu(p|\psi)$.



Dataset ($\mathcal{D}$)

| $A^1$ | $A^2$ | $\theta$ |
|---|---|---|
| $A_1^1$ | *unknown* | $\theta_1$ |
| *unknown* | $A_1^2$ | $\theta_2$ |
| $A_2^1$ | $A_2^2$ | $\theta_1$ |
| $A_1^1$ | $A_1^2$ | *unknown*= $\{\theta_1, \theta_2\}$ |
| $A_2^1$ | $A_2^2$ | $\theta_2$ |
| *unknown* | $A_1^2$ | $\theta_1$ |
| ⋮ | ⋮ | ⋮ |

Dataset ($\mathcal{D}_{CM}$)

| $A^1$ | $A^2$ | $\theta$ |
|---|---|---|
| $A_1^1$ | $A_1^2$ | *unknown*= $\{\theta_1, \theta_2\}$ |
| $A_2^1$ | $A_1^2$ | $\theta_2$ |
| $A_1^1$ | $A_2^2$ | $\theta_1$ |
| $A_2^1$ | $A_1^2$ | *unknown*= $\{\theta_1, \theta_2\}$ |
| $A_1^1$ | $A_2^2$ | $\theta_2$ |
| ⋮ | ⋮ | ⋮ |

Sub table 1: Complete cases $A^1$

| $A^1$ | $\theta$ |
|---|---|
| $A_1^1$ | $\{\theta_1, \theta_2\}$ |
| $A_2^1$ | $\theta_2$ |
| $A_1^1$ | $\theta_1$ |
| $A_2^1$ | $\{\theta_1, \theta_2\}$ |
| $A_1^1$ | $\theta_2$ |
| ⋮ | ⋮ |

Contingency table 1: $A^1$

| | $A_1^1$ | $A_2^1$ |
|---|---|---|
| $\theta_1$ | 20 | 30 |
| $\theta_2$ | 10 | 15 |
| $\{\theta_1, \theta_2\}$ | 5 | 2 |

Sub table 2: Complete cases $A^2$

| $A^2$ | $\theta$ |
|---|---|
| $A_1^2$ | $\{\theta_1, \theta_2\}$ |
| $A_1^2$ | $\theta_2$ |
| $A_2^2$ | $\theta_1$ |
| $A_1^2$ | $\{\theta_1, \theta_2\}$ |
| $A_2^2$ | $\theta_2$ |
| ⋮ | ⋮ |

Contingency table 2: $A^2$

| | $A_1^2$ | $A_2^2$ |
|---|---|---|
| $\theta_1$ | 30 | 40 |
| $\theta_2$ | 20 | 25 |
| $\{\theta_1, \theta_2\}$ | 6 | 9 |

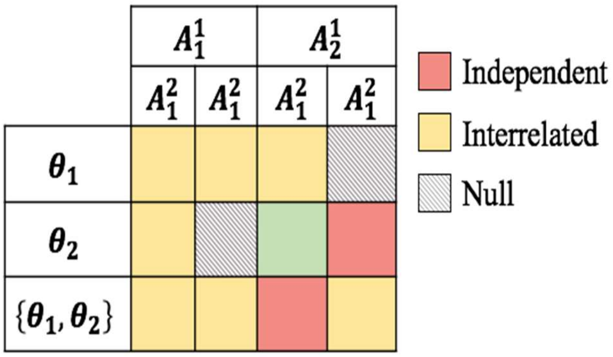
Sub table 3: Complete cases both $A^1$ and $A^2$

| $A^1$ | $A^2$ | $\theta$ |
|---|---|---|
| $A_1^1$ | $A_1^2$ | $\{\theta_1, \theta_2\}$ |
| $A_2^1$ | $A_1^2$ | $\theta_2$ |
| $A_1^1$ | $A_2^2$ | $\theta_1$ |
| $A_2^1$ | $A_1^2$ | $\{\theta_1, \theta_2\}$ |
| ⋮ | ⋮ | ⋮ |

Contingency table 3: both $A^1$ and $A^2$

| | $A_1^1$ | | $A_2^1$ | |
|---|---|---|---|---|
| | $A_1^2$ | $A_2^2$ | $A_1^2$ | $A_2^2$ |
| $\theta_1$ | 23 | 56 | 23 | 0 |
| $\theta_2$ | 56 | 0 | 67 | 12 |
| $\{\theta_1, \theta_2\}$ | 12 | 10 | 2 | 23 |

Figure 5. An example of subsets of complete data from an incomplete and ambiguous dataset



| | $A_1^1$ | | $A_2^1$ | |
|---|---|---|---|---|
| | $A_1^2$ | $A_1^2$ | $A_1^2$ | $A_1^2$ |
| $\theta_1$ | 🟨 | 🟨 | 🟨 | ▨ |
| $\theta_2$ | 🟨 | ▨ | 🟩 | 🟥 |
| $\{\theta_1, \theta_2\}$ | 🟨 | 🟨 | 🟥 | 🟨 |

🟥 Independent

🟨 Interrelated

▨ Null

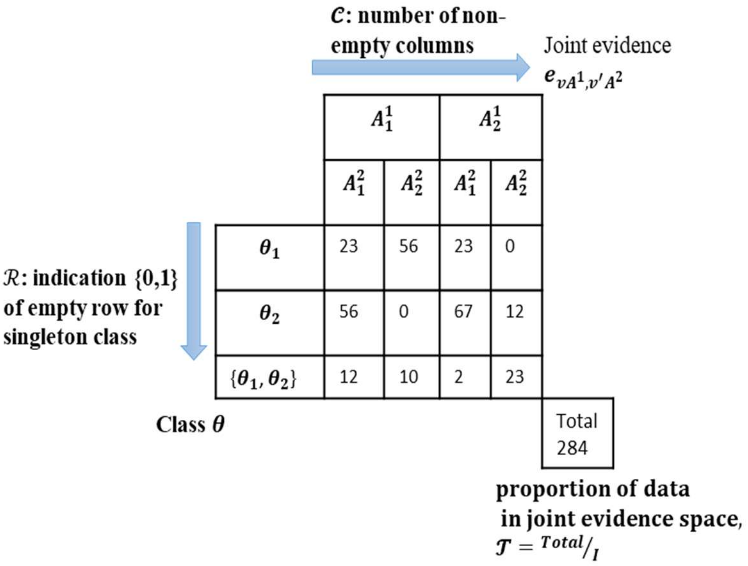
Figure 6. Interrelation

### 5.2.1.2 *Evidence Sparse Index:*

The columns of a contingency table of combined attributes represent joint pieces of evidence and rows represents classes $\theta$ such that $\theta \in P(\Theta)$. Suppose, there are two attributes $A^1$ and $A^2$. The contingency table for both attributes is shown in Figure 7. Each column of the contingency table represents joint pieces of evidence from attribute $A^1$ and $A^2$, for example, the first column represents joint evidence $\{A_1^1, A_1^2\}$. Each row of the contingency table represents the subset of the power set $P(\Theta) = \{\{\theta_1\}, \{\theta_2\}, \Theta = \{\theta_1, \theta_2\}\}$. Each cell of the contingency table represents the number of instances for a joint piece of evidence for a given class $\theta$.

Ideally, all data points (instances) for a piece of evidence are required in one of the cells to point correctly to a class. The reliability of evidence generated from a contingency table depends on its ability to point correctly to a class $\theta$. A good piece of evidence in an attribute would have a high density (or most of its samples) for a specific class. An empty column (joint evidence) indicates zero instance for any class. Real-world data usually does not have instances for all joint pieces of evidence in a dataset, which results in empty columns.

A dataset could be imbalanced or balanced due to a very small or almost equal number of instances for each class (or singleton subset). The zero number of instances in a row for a class is not sufficient to generate joint probability mass of multiple pieces of evidence in combined attributes. Furthermore, the proportion of the number of instances in a contingency table of joint pieces of evidence compared to the total number of instances ($I$) in a dataset $\mathcal{D}$ reflects the amount of statistical evidence. Suppose, the contingency table of two attributes $A^1$ and $A^2$ shown in Figure 7 has 284 available instances out of 2000 total number of instances in a dataset $\mathcal{D}$. The total number of instances or rows in a dataset is denoted by $I$. It shows that 14.2% of instances are in joint evidence space, and the remaining are missing. It suggests that two or more than two incomplete and ambiguous attributes can be combined by conjunctive MAKER when their contingency table does not have a large number of empty columns due to absence of the certain joint pieces of evidence for all classes, there are no empty rows due to absence of instances for a class or a singleton subset, and a sufficient number of instances exist in joint evidence space.


$\mathcal{C}$: number of non-empty columns
Joint evidence $e_{vA^1,v'A^2}$
$\mathcal{R}$: indication {0,1} of empty row for singleton class
Class $\theta$
proportion of data in joint evidence space, $\mathcal{T} = {}^{Total}/_{I}$

| | $A_1^1$ | | $A_2^1$ | |
|---|---|---|---|---|
| | $A_1^2$ | $A_2^2$ | $A_1^2$ | $A_2^2$ |
| $\theta_1$ | 23 | 56 | 23 | 0 |
| $\theta_2$ | 56 | 0 | 67 | 12 |
| $\{\theta_1, \theta_2\}$ | 12 | 10 | 2 | 23 |
| | | | | Total<br>284 |

Figure 7. Sparse index for contingency table of joint pieces of evidence in $A^1$ and $A^2$

These conditions are combined together by sparse index. It is denoted by $\mathcal{S}$. In the sparse index, the proportion of the number of non-empty columns for joint evidence denoted by $\mathcal{C}$; the indication of zero or non-zero instances for singleton subsets in the power set $\{\{\theta_1\}, \ldots \{\theta_z\}, \ldots, \{\theta_Z\}\}$, $z \in \{1, \ldots, z, \ldots, Z\}$ denoted by $\mathcal{R}$; and the proportion of data in the joint contingency table denoted by $\mathcal{T}$. The value of $\mathcal{R} = 0$ if, one of the singleton subsets represented by the row in contingency table has zero number of samples, otherwise $\mathcal{R} = 1$. The sparse index is given by:

$$\mathcal{S} = \frac{3\mathcal{CRT}}{\mathcal{CT} + \mathcal{RT} + \mathcal{CR}} \tag{16}$$

In Equation (16), $\mathcal{C}\epsilon[0,1]$, $\mathcal{R}\epsilon\{0,1\}$ and $\mathcal{T}\epsilon[0,1]$. The sparse index has values between zero and one, $\mathcal{S}\epsilon$ $[0,1]$. The sparse index is zero if any row except of missing labels $(\Theta)$ has zero instances and if all the columns are empty. The value of $\mathcal{T}$ is 1 if the input data has no missing values, $\mathcal{D}_{CM} = \mathcal{D}$.

### 5.2.2 Combine and Transform Interrelated Categorical Attributes

#### *5.2.2.1 Combine Evidences in Two Attributes*

The normalised joint probability mass for combined pieces of evidence in categorical attributes $A^1$ and $A^2$ is denoted by $\widehat{m}_{\theta,v1,v_2'}$, where $v \in V_1$ and $v' \in V_2$ are $v^{th}$ and $v'^{th}$ evidence in $A^1$ and $A^2$, respectively. The joint probability mass $(\widehat{m}_{\theta,v1,v_2'})$ that class $\theta$ is supported by evidence $e_{v,A^1}$ and $e_{v',A^2}$ is given by:

$$\widehat{m}_{\theta,v1,v'2} = \begin{cases} 0 & \theta = \emptyset \\ \frac{m_{\theta,v1,v_2'}}{\sum_{\theta\in P(\Theta)} m_{\theta,v1,v_2'} + m_{P(\Theta),v1,v_2'}} & \forall\theta \in P(\Theta), \theta \neq \emptyset \end{cases} \tag{17a}$$

$$m_{\theta,v1,v'2} = \left[(1 - r_{v',2})m_{\theta,v,1} + (1 - r_{v,1})\, m_{\theta,v',2}\right] + \sum\nolimits_{h1\cap h2=\theta} \gamma_{h1,h2,v1,v_2'}\, \psi_{h1,h2,v1,v_2'}\, m_{h1,v,1}\, m_{h2,v',2} \tag{17b}$$

The residual support $(m_{P(\Theta),v1,v_2'})$ in Equation (17a) is earmarked to the power set as given by:

$$m_{P(\Theta),v1,v_2'} = m_{\Theta,v,1}\, m_{\Theta,v',2} \tag{17c}$$

The Equation (17a) is the normalised joint probability mass obtained from Equations (17b) and (17c). In Equation (17b), $m_{\theta,v,1}$ and $m_{\theta,v',2}$ are the probability masses of single evidence in attributes $A^1$ and $A^2$, respectively. It is obtained by Equation (7), which consists of weight of evidence $(w_{\theta,v,q})$ and basic probability $(p_{\theta,v,q})$. The reliability $(r_{\theta,v,q})$ of $v^{th}$ evidence in $q^{th}$ categorical attribute pointing to a class $\theta$ is obtained by Equation (7). The overall reliability of evidence $(r_{v,q})$ is the sum of the products of $r_{\theta,v,q}$ and $p_{\theta,v,q}$ for all $\theta \in P(\Theta)$. The reliability of evidence of a $v^{th}$ evidence in a $q^{th}$ evidence is given by:

$$r_{v,q} = \sum\nolimits_{\theta\in P(\Theta)} r_{\theta,v,q}\; p_{\theta,v,q} \tag{18}$$

The parameter $\gamma_{h1,h2,v1,v_2'}$ is called the reliability ratio. It is the ratio of the joint reliability of the two pieces of evidence and the product of their individual reliabilities. The weight of evidence and reliability ratio can be trained by data-driven optimisation. The initial weight of evidence can be assumed equal to reliability $(w_{\theta,v,q} = r_{\theta,v,q})$.

#### *5.2.2.2 Combine Evidences in Multiple Attributes*

This framework can combine evidences in multiple attributes. The multiple attributes are combined in an iterative manner, i.e. first, the evidences in two individual attributes are combined, then these two combined attributes are combined with the third attribute. This process continues until a group of interrelated attributes are combined. This group of interrelated attributes would have a size greater than or equal to two. The technique to find interrelated attributes based on interrelation index and sparse index is demonstrated in Section 5.2.1. Suppose we want to combine $Q^{'}$ categorical attributes $A^1, \dots, A^q, \dots, A^{Q^{'}}$ each with $V_q$ pieces of evidences. Single and joint probability are obtained from Equation (5), $p_{\theta,v,1},\dots, p_{\theta,v,q},\dots, p_{\theta,v,Q^{'}}$ and $p_{\theta,v1,v2},\dots, p_{\theta,v1,\dots,vq},\dots, p_{\theta,v1,\dots,VQ^{'}}$ such that $v \in \{1, \dots, V_q\}$, $\forall\theta \in P(\Theta)$. The interrelation between evidence, i.e. $\psi_{\theta,v1,v2},\dots, \psi_{\theta,v1,\dots,vq},\dots, \psi_{\theta,v1,\dots,VQ^{'}}$ , is obtained from Equations (14a) and (14b). Single probability mass, i.e. $m_{\theta,v,1}, \dots, m_{\theta,v,q}, \dots, m_{\theta,v,Q^{'}}$ is obtained from Equation (6). Combined probability mass is calculated iteratively once single probability, joint probability, and interrelation are calculated from the data. The degree of the combined support $m_{\theta,v1,v2}$ for joint evidence $e_{v,1}$ and $e_{v,2}$ for attribute $A^1$ and $A^2$, respectively pointing towards proposition $\theta$, $\theta \in P(\Theta)$ is given by:

$$m_{\theta,v1,v2} = \left[(1 - r_{v,2})m_{\theta,v,1} + (1 - r_{v,1})\, m_{\theta,v,2}\right] + \textstyle\sum_{h1 \cap h2 = \theta} \gamma_{h1,h2,v1,v2}\, \psi_{h1,h2,v1,v2}\, m_{h1,v,1}\, m_{h2,v,2} \tag{19}$$

The degree of the combined support from evidence $e_{v1,v2}$ obtained from Equation (20) and evidence $e_{v,3}$ in attribute $A^3$ pointing to preposition $\theta$ is given by:

$$m_{\theta,v1,v2,v3} = \left[(1 - r_{v,3})m_{\theta,v1,v2} + m_{P(\theta),v1,v2}\, m_{\theta,v,3}\right] + \textstyle\sum_{h1 \cap h2 = \theta} \gamma_{h1,h2,v1,v2,v3}\, \psi_{h1,h2,v1,v2,v3}\, m_{h1,v1,v2}\, m_{h2,v,3} \tag{20}$$

The above process is repeated until all $Q^{\prime}$ number of attributes are combined to generate the degree of the combined support $m_{\theta,v1,\ldots,VQ^{\prime}}$. It is then normalised by using Equation (17a) to obtain the combined probability mass $\widehat{m}_{\theta,v1,\ldots,VQ^{\prime}}$. The combined probability mass $m_{\theta,v1,\ldots,VQ^{\prime}}$ pointing to proposition $\theta$ is given by:

$$m_{\theta,v1,\ldots,VQ^{\prime}} = \left[(1 - r_{v,Q^{\prime}})\, m_{\theta,v1,\ldots,VQ^{\prime}-1} + m_{P(\theta),v1,\ldots,vQ^{\prime}-1} m_{\theta,v,Q^{\prime}}\right] + \\ \textstyle\sum_{h1 \cap h2 = \theta} \gamma_{h1,h2,v1,\ldots,vQ^{\prime}}\, \psi_{h1,h2,v1,\ldots,vQ^{\prime}}\, m_{h1,v1,\ldots,VQ^{\prime}-1}\, m_{h2,v,Q^{\prime}} \tag{21a}$$

The combined probability mass left for the power set $m_{P(\Theta),v1,\ldots,VQ^{\prime}}$ is given by:

$$m_{P(\Theta),v1,\ldots,VQ^{\prime}} = m_{P(\Theta),v1,\ldots,VQ^{\prime}-1}\, m_{P(\Theta),v,Q^{\prime}} \tag{21b}$$

The combined probability $\widehat{m}_{\theta,v1,\ldots,VQ^{\prime}}$ after normalisation of combined probability mass is:

$$\widehat{m}_{\theta,v1,\ldots,VQ^{\prime}} = \frac{m_{\theta,v1,\ldots,VQ^{\prime}}}{\sum_{g \in P(\Theta)} m_{g,v1,\ldots,VQ^{\prime}} + m_{P(\Theta),v1,\ldots,VQ^{\prime}}}\ \forall\theta \in P(\theta) \text{ and } \widehat{m}_{P(\Theta),v1,\ldots,VQ^{\prime}} = \frac{m_{P(\Theta),v1,\ldots,VQ^{\prime}}}{\sum_{g \in P(\Theta)} m_{g,v1,\ldots,VQ^{\prime}} + m_{P(\Theta),v1,\ldots,VQ^{\prime}}} \tag{21c}$$

The MAKER framework has two types of parameters – weight and reliability ratio. The parameters are optimised by maximising the likelihood of the true state.

$$\text{Minimise:} f(parameter) = \frac{1}{2I} \textstyle\sum_{i=1}^{I} \sum_{\theta \in P(\theta)} \left(\widehat{m} - \widehat{m}(parameters)\right)^2 \tag{22}$$

where, $parameters = parameters \begin{pmatrix} \text{Weight: } w_{\theta,v,1},\ldots,w_{\theta,v,q},\ldots,w_{\theta,v,Q},w_{\theta,v1,v2},\ldots,w_{\theta,v1,..,vQ} \\ \text{Reliability ratios: } \gamma_{h1,h2,v1,v2},\ldots,\gamma_{h1,h2,v1,\ldots,vQ} \end{pmatrix}$

The algorithmic steps to combine multiple categorical attribute by conjunctive MAKER are shown in Table 17, Appendix A. The conjunctive MAKER algorithm to combine and transform multiple incomplete and ambiguous categorical attribute is denoted as C-MAKER.

### 5.3 Computational complexity of I-MAKER and C-MAKER

The main computation cost of MAKER is in training (optimization) of the parameters and repetition of this process to conduct *k*-fold regularisation to suppress data leakage, overfitting, and smoothing to overcome the problem of under-sampling. The parameters include the weight and reliability of each evidence in both I-MAKER and C-MAKER and an additional parameter: the reliability ratio in C-MAKER. A fast pre-processing of incomplete and ambiguous categorical attributes can be achieved by assuming that the evidence with the highest reliability could have relatively high importance compared to other evidence. It suggests that the training time can be reduced significantly by concluding that the reliability of evidence is equivalent to its weight.

The iterations for $k$-fold regularisation to estimate the probability mass of evidence in the attributes in a dataset are independent of each other. Therefore, a feasible runtime for pre-processing a dataset by I-MAKER and C-MAKER can be achieved by parallel implementations. The execution time of the tasks in an algorithm is defined by Big O notation, which stands for "order of magnitude". Both algorithms, for I-MAKER shown in Table 16 and C-MAKER shown in Table 17, have a non-linear computational complexity that can be roughly estimated as $O(Q) \times O(K \times O(training))$ and $O(cardinality\ of\ \overline{\overline{\Omega}}) \times O(K \times O(training))$, respectively. In estimated time complexity, $O(training)$ is the complexity of optimization, or training of parameters, in the MAKER, $Q$ is the number of categorical attributes, $K$ is the number of folds for the regularisation, and $cardinality\ of\ \overline{\overline{\Omega}}$ represents the number of feasible combinations of attributes in the set $\overline{\overline{\Omega}}$. The first Big O, $O(Q)$ and $O(cardinality\ of\ \overline{\overline{\Omega}})$ for I-MAKER and C-MAKER, respectively, represents the computational cost of the first *for-loop* in the algorithm. Similarly, the second Big O, $O(K \times O(training))$, represents the computational cost of the second *for-loop* in the algorithm. There is a one-time computational cost of training and regularisation for estimation of the probability mass, $O(K \times O(training))$, before the deployment of an AI/ML model. The newly arrived data will be transformed based on previously estimated probability mass. The probability mass can be updated regularly when an adequate amount of new data is available.

## 6. Numerical Example

The numerical example in this section demonstrates the steps to combine two interrelated attributes in early asthma signs and symptoms dataset. The two attributes, the physical exercise (E) and peak expiratory rate (PE) have the highest number of interrelated evidences with the highest joint probability towards a specific class in $P(\Theta)$ (Power set shown in Equation (2). The concluded results of interrelation test on asthma data is shown in Table 27 in Appendix C. The asthma data obtained from the National Health Services (NHS) in the UK has four attributes. This dataset will be discussed further in the case study in Section 7. The dataset has 4827 cases. A small percentage of samples are kept aside for validation and the rest of the samples are split into five cross-validation sets to obtain the combined probability mass for each set. The cross-validation data is partitioned into 3474 samples for three sets and 3475 samples for the other two sets. The frame of discernment in this example is $\theta = \{ER, N\}$, where $ER$ represents an early sign and $N$ represents no sign. The power set is $P(\Theta) = \{ER, N, unknown\}$ where $unknown = \Theta = \{ER, N\}$. The sets $\{E1, E2, E3\}$ and $\{PER1, PER2, PER3\}$ contain evidences (which can also be called referential values or categories) in the attributes E and PE, respectively. Following are the steps to fuse two interrelated attributes by joining the evidences to obtain the combined probability mass for numerical data transformation. The following steps are shown for one cross-validation set. These sets are performed for all five cross-validation sets of both attributes. The average combined probability mass of all five cross-validation sets is used for data transformation of joint evidences. The set of joint evidences in both attributes can be written as $\{(E1, PER1), (E2, PER1), (E3, PER1), (E1, PER2), (E2, PER2), (E3, PER2), (E1, PER3), (E2, PER3), (E3, PER3)\}$.

*STEP 1: Contingency table*

The contingency tables containing the number of samples for single evidence in E and PE and joint evidences in both attributes are shown in Table 2 and Table 3. In Table 3, some of the joint evidences have zero samples due to the unavailability of those cases in the dataset. The likelihood and basic probability are obtained from the contingency tables.

Table 2: Contingency table of an attribute

| ***Physical exercise (E)*** | E1 | E2 | E3 |
|---|---|---|---|
| ***Early Diagnosis*** | | | |
| UNKNOWN ($\theta$) | 12 | 0 | 1 |

| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 |
|---|---|---|---|
| ***Early Diagnosis*** | | | |
| UNKNOWN ($\theta$) | 4 | 8 | 1 |

| | | | |
|---|---|---|---|
| ER | 333 | 245 | 7 |
| N | 1127 | 932 | 31 |

| | | | |
|---|---|---|---|
| ER | 483 | 88 | 14 |
| N | 19 | 2065 | 6 |

Table 3: Contingency table of the two attributes

| ***Physical exercise (E)*** | E1 | | | E2 | | | E3 | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 |
| ***Early Diagnosis*** | | | | | | | | | |
| UNKNOWN ($\Theta$) | 4 | 7 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| ER | 276 | 47 | 10 | 202 | 40 | 3 | 5 | 1 | 1 |
| N | 16 | 1107 | 4 | 3 | 927 | 2 | 0 | 31 | 0 |

*STEP 2: Likelihood of singleton and joint evidences*

The likelihood ($L_{\theta,v,q}$) is obtained from contingency tables by performing a division operation in each row by its sum. The likelihood of singleton and joint evidences in E and PE is shown in Table 18 and 19 in Appendix B.

*STEP 3: Basic probability of singleton and joint evidences*

The basic probability is obtained from the likelihood. The basic probability for singleton and joint evidences in both the attributes is obtained from Equation (5). The following Tables 4 and 5 shows the singleton and joint basic probability of evidences pointing towards $P(\Theta)$ in attributes E and PE, respectively.

Table 4: Basic probability of E and PE

| ***Physical exercise (E)*** | E1 | E2 | E3 |
|---|---|---|---|
| ***Early Diagnosis*** | | | |
| $\Theta$ | 0.440 | 0.064 | 0.706 |
| ER | 0.292 | 0.446 | 0.110 |
| N | 0.268 | 0.490 | 0.185 |

| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 |
|---|---|---|---|
| ***Early Diagnosis*** | | | |
| $\Theta$ | 0.328 | 0.319 | 0.701 |
| ER | 0.665 | 0.086 | 0.268 |
| N | 0.007 | 0.596 | 0.031 |

Table 5: Joint probability of evidences in E and PE

| ***Physical exercise (E)*** | E1 | | | E2 | | | E3 | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 |
| ***Early Diagnosis*** | | | | | | | | | |
| $\Theta$ | 0.391 | 0.469 | 0.802 | 0.000 | 0.000 | 0.000 | 0.000 | 0.823 | 0.000 |
| ER | 0.599 | 0.070 | 0.178 | 0.996 | 0.134 | 0.843 | 1.000 | 0.018 | 1.000 |
| N | 0.010 | 0.461 | 0.020 | 0.004 | 0.866 | 0.157 | 0.000 | 0.159 | 0.000 |

*STEP 4: Interrelation index*

After the single and joint basic probability have been acquired, the Interrelation index between two pieces of evidence is calculated through Equations (14a) and (14b). Table 6 shows the interrelation between each two pieces of evidence in E and PE.

Table 6: Interrelation index

| ***Physical exercise (E)*** | E1 | | | E2 | | | E3 | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 |
| ***Early Diagnosis*** | | | | | | | | | |
| $\Theta$ | 2.71 | 3.34 | 2.60 | 0.00 | 0.00 | 0.00 | 0.00 | 3.66 | 0.00 |
| ER | 3.09 | 2.80 | 2.28 | 3.36 | 3.50 | 7.05 | 13.72 | 1.95 | 34.04 |
| N | 5.26 | 2.89 | 2.40 | 1.23 | 2.97 | 10.37 | 0.00 | 1.44 | 0.00 |

*STEP 5: Reliability of the evidence*

The reliability of the evidence is defined in Section 5.1. It is the ability of evidence (both single and joint) to point correctly to a class in $P(\Theta) = \{ER, N, unknown\}$. The reliability of the evidence pointing to a class is obtained from Equation (7) and the overall reliability of the evidence weighted by basic probability is obtained from Equation (18). The reliability of single and joint evidence is shown in Table 20 and 21 in Appendix B, respectively.

*STEP 6: Set initial values for parameters*

The weight of the evidence and the reliability ratio are two parameters in the MAKER framework. Initially, the weight of the evidence is assumed equal to the reliability of the evidence obtained in the previous step. All initial values of the reliability ratio are assumed to be equal to 1. The initial values of the weight of the evidences is shown in Table 20 and 21 in Appendix B. The initial reliability ratio of the evidences are shown in Table 22 in Appendix B.

*STEP 7: Initial probability mass of single attributes (uncombine)*

The probability mass is obtained by Equation (6). It is the multiplication of the basic probability and the weight of the evidence. The probability mass for uncombine evidences in the attributes E and PE is obtained first, then the combined probability mass is obtained in the next step.

Table 7: Initial probability mass for attribute E and PE

| ***Physical exercise (E)*** | E1 | E2 | E3 |
|---|---|---|---|
| ***Early Diagnosis*** | | | |
| UNKNOWN ($\Theta$) | 0.011 | 0.000 | 0.091 |
| ER | 0.320 | 0.252 | 0.114 |
| N | 0.669 | 0.748 | 0.795 |

| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 |
|---|---|---|---|
| ***Early Diagnosis*** | | | |
| $\Theta$ | 0.005 | 0.000 | 0.131 |
| ER | 0.995 | 0.006 | 0.833 |
| N | 0.000 | 0.994 | 0.037 |

*STEP 8: Combined probability mass before training*

The combined probability mass $(\widehat{m}_{\theta,v1,v'2})$ is obtained from the Equations (17a) to (17c). Equations (17b) and (17c) calculate probability mass and residual support, respectively. The combined probability mass is then normalised by the sum of the probability mass and residual support for the joint evidence by Equation (17a). The combined probability mass before training of parameters is shown in Table 8. Equations (21a) to (21b) are used if evidences in more than two attributes are combined.

Table 8: Estimated combined probability: before training

| ***Physical exercise (E)*** | E1 | | | E2 | | | E3 | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 |
| ***Early Diagnosis*** | | | | | | | | | |
| *Θ* | 0.146 | 0.198 | 0.361 | 0.000 | 0.000 | 0.000 | 0.000 | 0.439 | 0.000 |
| ER | 0.672 | 0.184 | 0.411 | 0.890 | 0.311 | 0.459 | 1.000 | 0.109 | 1.000 |
| N | 0.182 | 0.618 | 0.227 | 0.110 | 0.689 | 0.541 | 0.000 | 0.453 | 0.000 |

*STEP 9: Combined probability mass after training*

Table 9 shows the combined probability after training of the weight and residual support of the evidence. Table 10 shows the probability mass of evidences in individual attributes. The value of trained parameters is shown in Tables 23 to 25 in Appendix B. The joint probability of some of the pieces of evidence in Table 5 is zero, due to missing values in the contingency table (Table 3). After optimisation, the combined probability mass of support for each state is predicted.

Table 9: Estimated combined probability: after training

| ***Physical exercise (E)*** | E1 | | | E2 | | | E3 | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 |
| ***Early Diagnosis*** | | | | | | | | | |
| *Θ* | 0.045 | 0.030 | 0.343 | 0.011 | 0.006 | 0.001 | 0.010 | 0.075 | 0.007 |
| ER | 0.913 | 0.049 | 0.434 | 0.928 | 0.031 | 0.358 | 0.987 | 0.087 | 0.991 |
| N | 0.041 | 0.921 | 0.224 | 0.060 | 0.963 | 0.641 | 0.003 | 0.839 | 0.002 |

Table 10: Updated probability of E and PE

| ***Physical exercise (E)*** | E1 | E2 | E3 |
|---|---|---|---|
| ***Early Diagnosis*** | | | |
| *Θ* | 0.012 | 0.000 | 0.079 |
| ER | 0.343 | 0.369 | 0.230 |
| N | 0.645 | 0.631 | 0.691 |

| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 |
|---|---|---|---|
| ***Early Diagnosis*** | | | |
| *Θ* | 0.005 | 0.000 | 0.197 |
| ER | 0.995 | 0.006 | 0.738 |
| N | 0.000 | 0.994 | 0.065 |

*STEP 10: Average probability mass*

Step 1 to Step 9 in this example, demonstrate the methodology used to obtain the combined probability mass from data in the first fold. The average of combined and single probability mass of evidences from all five folds in this numerical example is used for data transformation of joint pieces of evidence in E and PE is shown in following Tables 11 and 12.

Table 11: Average estimated combined probability

| ***Physical exercise (E)*** | E1 | | | E2 | | | E3 | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 |
| ***Early Diagnosis*** | | | | | | | | | |
| *Θ* | 0.035 | 0.08 | 0.356 | 0.021 | 0.07 | 0.008 | 0.01 | 0.075 | 0.009 |
| ER | 0.914 | 0.08 | 0.42 | 0.91 | 0.03 | 0.36 | 0.987 | 0.087 | 0.99 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| N | 0.041 | 0.84 | 0.224 | 0.07 | 0.9 | 0.631 | 0.003 | 0.839 | 0.008 |

Table 12: Average updated probability of E and PE

| ***Physical exercise (E)*** | E1 | E2 | E3 | | ***Peak expiration (PE)*** | PER1 | PER2 | PER3 |
|---|---|---|---|---|---|---|---|---|
| ***Early Diagnosis*** | | | | | ***Early Diagnosis*** | | | |
| $\Theta$ | 0.012 | 0.000 | 0.079 | | $\Theta$ | 0.038 | 0.089 | 0.177 |
| ER | 0.353 | 0.349 | 0.330 | | ER | 0.897 | 0.013 | 0.738 |
| N | 0.635 | 0.651 | 0.600 | | N | 0.064 | 0.900 | 0.075 |

*STEP 11: Data fusion and transformation*

The combined probability mass obtained from the MAKER framework is used to fuse and transform two or more attributes into $Z$ or $Z+1$ dimensional numerical features. Each feature represents a possible outcome or state $\theta \in P(\Theta)$. Table 13 demonstrates the example of fused and transformed data in attributes E and PE into 3-dimensional numerical feature using the probability mass in Tables 11 and 12. The first-dimension feature labelled as unknown covers uncertainty due to ambiguity in the dataset. The second and third dimension represent early signs of asthma and no signs of asthma, respectively. In Table 13, if only one asthma symptom is available (other symptom is missing) then probability mass of single evidence is referred from Table 12, if both symptoms are available then Table 11 is used for data transformation. The example of the data transformation of all the attributes in the asthma data is demonstrated in Table 26 in Appendix B. The interrelation tests demonstrated in Section 5.2.1 for both datasets are presented in Tables 27 and 28 in Appendix C.

Table 13: An example of data fusion and interpretable transformation by C-MAKER

| | ***Attribute 1*** | ***Attribute 2*** | | ***Transformed data*** | | |
|---|---|---|---|---|---|---|
| # data points | ***Physical exercise (E)*** | ***Peak expiration (PE)*** | | ***Unknown ($\Theta$)*** | ***Early sign (ER)*** | ***No signs (N)*** |
| 1 | E1 | PER1 | | 0.035 | 0.914 | 0.041 |
| 2 | | PER3 | | 0.177 | 0.738 | 0.075 |
| 3 | E2 | PER1 | ➡ | 0.021 | 0.91 | 0.07 |
| 4 | E2 | | | 0.00 | 0.349 | 0.651 |
| 5 | E2 | PER2 | | 0.07 | 0.03 | 0.90 |
| 6 | E3 | PER2 | | 0.075 | 0.087 | 0.839 |
| 7 | | PER3 | | 0.177 | 0.738 | 0.075 |
| ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ |

## 7. Case Study on Asthma and Mortgage Loan

### 7.1 Approach for Performance and Sensitivity Analysis

Two real-world data were selected to demonstrate the performance and uncertainty management of MAKER for the interpretable transformation of incomplete and ambiguous categorical attributes. The proposed methodology applies to individual or multiple categorical attributes. The methodology for an individual attribute is called I-MAKER and for multiple attributes is called C-MAKER. The application of both methods is demonstrated in healthcare and finance data. The healthcare data is about early asthma symptoms in children. It is incomplete and ambiguous data (missing values in both the input and output attributes). The finance data is about mortgage loans. It is incomplete and unambiguous data; only values in the input attributes are missing. Both domains are highly regulated. Decision-making by manual, semi-automated or solely automated system has high stakes in both domains. Therefore, data pre-processing steps require awareness about existing uncertainty in the data and flexibility to include expert judgment.

The performance of I-MAKER and C-MAKER were compared with missForest, MICE, EM, and KNN. These four techniques are the most commonly used for data imputation. Studies have (Stekhoven & Bühlmann, 2012) (Waljee A. , et al., 2013) shown that among all methods, missForest is the most efficient, in terms of accuracy and imputation error of prediction models. Three types of AI models- ANN, decision tree, and BRB were selected for performance and sensitivity analyses of asthma and loan data for three different missing data scenarios. An ANN is a deep learning model, a decision tree is a tree-based model, and a BRB is a rule-based model. BRB is an extension from the IF THEN rule-based system that enables the application of belief distribution in presenting the relationship between predictors and outcomes in a transparent and interpretable way (Yang, Liu, Wang, Sii, & Wang, 2006). It can have hierarchical and non- hierarchical structure.

In scenario I, all six methods (I-MAKER and C-MAKER with four other data imputation methods) were tested using original data with the three ML algorithms. In scenario II, 20% of the data were randomly removed from the original dataset by adding 20% auxiliary missingness to the original data. Therefore, the total proportion of missingness in scenario II was the original data missing $\% + 20\%$. Similarly, in scenario III, 35% of the data were randomly removed from the original dataset. Then, data was split into two parts: training (80%) and validation (20%). Both the training and validation sets were stratified on the output attribute to balance the distribution of all classes, $\theta \in P(\Theta)$. The validation data samples were not part of the training data; they were solely used to assess the performance and sensitivity of data imputed by six different imputation methods on neural network, decision tree, and belief-rule-base models. Both I-MAKER and C-MAKER were trained separately for each five-fold stratified cross-validation sets to randomise the less frequent evidence. The average probability mass of evidence (or categories) obtained from the five training folds was used to transform the categorical attributes. In I-MAKER and C-MAKER, uncertainty due to missingness in the output attribute is considered in the estimation of probability mass for all possible states in a power set. Other data imputation methods cannot consider such uncertainty; therefore, such samples are thoroughly ignored in these methods.

All nominal categorical attributes (no intrinsic order) were transformed into dummy variables $\{0,1\}$, and all ordinal categorical attributes (intrinsic order) were transformed into unique integers $\{1,2,3, \dots\}$ using label encoding technique. The missing values imputed in a categorical attribute by missForest, MICE, EM, and KNN were rounded to the nearest integer value that belonged to the set of integers used to label that attribute. For instance, in numerical example in Section 6, the attribute peak expiration rate (PER) in the asthma dataset has three levels, from lowest to highest $\{\text{PER1}, \text{PER2}, \text{PER3}\}$. These three levels can be transformed into a numerical attribute using the label encoding technique $\{\text{PER1} = 1, \text{PER2} = 2, \text{PER3} = 3\}$. Suppose, a missing value $2.3$ is predicted by the data imputation technique missForest, then, this value would be rounded to the closest integer (discrete) $2$, to point clearly to the category $\text{PER2} = 2$. The AI/ML models can process both discrete and continuous data. However, the approximation of a continuous number to a discrete number for a categorical attribute would be a vital step if a model-agnostic or model-specific method were used to explain the decisions by a black-box model. The explainability and uncertainty management of MAKER will be discussed in Section 7.4. Depending on the type of AI/ML model, the imputed input data can be encoded into the most preferred format, such as one-hot encoding for ANN, label encoding for decision-tree, and belief-distribution for BRB.

### 7.2 Results: Early Asthma Symptoms

The early asthma symptoms data in children was obtained from NHS Digital in the UK. Children’s asthma is the third highest medical condition in the UK and the most common reason for urgent admission to hospitals. The dataset was utilised to detect early asthma signs and symptoms in children. It has four categorical attributes: sleep disturbance

(S), a nocturnal symptom; the existence of daytime (DT) symptoms; peak expiratory rate (PER), the measure of the maximum amount of air that can be exhaled from the lungs; and triggers by physical exercise or activity (E). The dataset has 4827 cases. The missing values exist in all four input attributes for asthma symptoms and output attribute for asthma diagnosis. In total, 10.99% of the data is missing. The record of asthma diagnosis has two outcomes (decisions); the set of outcomes can be written as $\Theta = \{early\ sign, no\ sign\}$. The power set of all possible states is $P(\Theta) = \{\{early\ sign\}, \{no\ sign\}, \Theta = \{early\ sign, no\ sign\}\}$. The probability mass for all evidence in the data by I-MAKER and C-MAKER was estimated for all states in $P(\Theta)$. The proportion of missingness in each attribute can be seen in Figure 8.

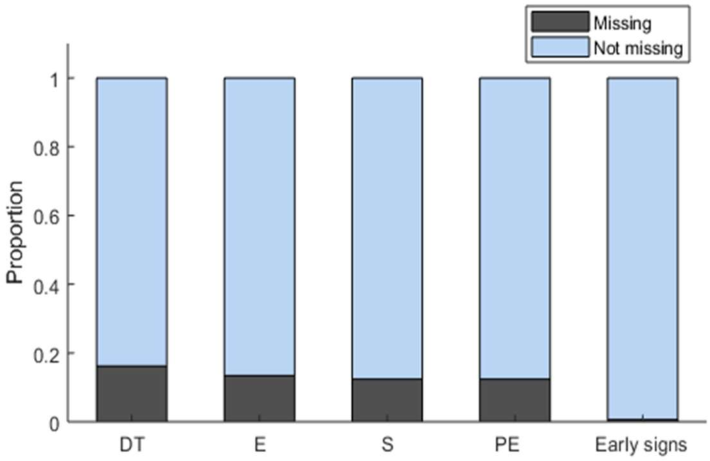


Figure 8. The proportion of missing values in early asthma symptoms data

The performance and sensitivity of I-MAKER and C-MAKER compared to other data imputation methods for incomplete and ambiguous asthma data was analysed through three different scenarios for three AI models. The hyper-parameters of ANN, decision tree, the BRB structure for individual attributes transformed by I-MAKER, and combined by C-MAKER is shown in Table 29, Table 31, Figure 15, and Figure 16, respectively. The first layer $(L_1)$ in an ANN is the input layer. In Table 29, it can be seen that the input data transformed by C-MAKER and I-MAKER in the third scenario has only 9 and 18 transformed features, respectively, compared to 14 features from other imputation methods for ANN models. In the asthma data, two sets of attributes $\{E, PE\}$ and $\{S, DT\}$ were combined by the C-MAKER rule. Table 27 in Appendix C shows the proportion of interrelated evidence, joint probability, and sparse index. The set $\{E, PE\}$ has a high proportion of interrelated evidence, $\psi^p = 1.0$; a high joint probability, $\mu(p|\psi) = 0.79$; and a good sparse index, $\mathcal{S} = 0.909$. The set $\{S, DT\}$ has a proportion of interrelated evidence of $\psi^p = 0.916$, a joint probability of $\mu(p|\psi) = 0.58$, and a good sparse index of $\mathcal{S} = 0.915$.

The incompleteness of data was simulated by varying the proportion of missingness. Table 14 demonstrates the area under the ROC curve (AUC) score of ANN, decision tree, and BRB for missing data imputed by six different methods under three missing data scenarios: 10.99%, 30.99% (10.99%+20%), and 45.99% (10.99%+35%). The increase in missingness proportion in the data results in a loss of information. The best performance can be seen in the scenario I and the worst in scenario III. The missingness decreased the abstraction capacity of any learning algorithm; however, the inclusion of uncertainty in the data controls biases and errors. In scenario I, most of the performance of data transformed by I-MAKER and C-MAKER is very close to the performance of data imputed by missForest. For all three models under all three different scenarios, missForest performed better than MICE. The performance gap between I-MAKER and C-MAKER with other data imputation techniques increased as the missing proportion increased. In scenario II, I-MAKER and C-MAKER had relatively similar performances, and both performed considerably better than data imputation techniques. In scenario III, C-MAKER did not perform well when compared to I-MAKER for all three models, due to insufficient samples when fusing two or more pieces of evidence in the combined evidence space. The relatively common order of AUC from the highest for scenario III, which has 45.99% of data missing, was I-MAKER ≈ C-MAKER > missForest > MICE > EM > KNN.

Table 14: AUC Performance comparison for asthma symptoms data

| | Scenario I<br>Original data missing proportion = 10.99% | | | | | |
|---|---|---|---|---|---|---|
| | **I-MAKER** | **C-MAKER** | **missForest** | **MICE** | **EM** | **KNN** |
| **ANN** | 0.959 | 0.962 | 0.962 | 0.960 | 0.958 | 0.959 |
| **Decision tree** | 0.963 | 0.959 | 0.963 | 0.901 | 0.899 | 0.962 |
| **BRB** | 0.961 | 0.965 | 0.945 | 0.943 | 0.952 | 0.956 |
| **Scenario II**<br>**Missing proportional: original data % + 20% = 30.99%** | | | | | | |
| **ANN** | 0.830 | 0.834 | 0.829 | 0.804 | 0.791 | 0.782 |
| **Decision tree** | 0.816 | 0.813 | 0.780 | 0.740 | 0.618 | 0.795 |
| **BRB** | 0.868 | 0.838 | 0.820 | 0.800 | 0.766 | 0.770 |
| **Scenario III**<br>**Missing proportional: original data % + 35% = 45.99%** | | | | | | |
| **ANN** | 0.744 | 0.710 | 0.609 | 0.593 | 0.511 | 0.520 |
| **Decision tree** | 0.745 | 0.732 | 0.540 | 0.532 | 0.515 | 0.500 |
| **BRB** | 0.733 | 0.690 | 0.599 | 0.592 | 0.509 | 0.576 |

## 7.3 Results: Mortgage Loan

The mortgage loan data was obtained from Together Financial Services, a mortgage lending firm in the UK. The attributes of the loan data relate to affordability, unsecured loans, secured loans, bankruptcy and payday, debit and debit searches, credit score, loan criteria, property valuation, and property value. There are 18 attributes, named sequentially from $A1$ to $A18$. Only attributes related to credit score are quantitative; all other attributes are categorical in nature. The credit score is a continuous attribute. Missing values were filled with an extreme negative value (Saar-Tsechansky & Provost, 2007). The mortgage loan data is incomplete and unambiguous since it has missing values in the attributes of input data and no missing class labels. The percentage of missingness in each attribute can be seen in Figure 9. In a BRB model, the referential values of quantitative attributes are trained. The trained referential values of credit score for BRB model are $\{-99.0,0.0,256.31,410.56,600\}$; here -99 is the lower bound, which refers to an unknown credit score for a customer, and 600 is the upper bound of the credit score. The loan data consists of historical credit data from a credit bureau and features extracted from electronic loan applications. The columns in the dataset were aggregated using average, maximum, minimum, and sum operations. For example, data for the rule 'worst status of secured loans in the last 12 months' would represent the maximum of worst status columns for the current address, previous address, and linking address. Likewise, data for the rule 'number of bankruptcies in the last six years' is the sum of the columns for the number of satisfied and unsatisfied bankruptcies in the last six years. The explainable system based on BRB was developed to automate the mortgage loan application process (Sachan S. , Yang, Xu, Benavides, & Li, 2020). A detailed description of this data and the methodology for data treatment can be seen in the aforementioned paper. The loan dataset has 3498 cases. In total, 6.55% of the data is missing and only attributes $A1, A3, A10, A12, A13,$ and $A16$ have missing values. Attribute A16 for property valuation has the highest number of missing values ($\cong 72.06\%$). The set of decisions in loan data is $\Theta = \{\text{fund}, \text{reject}\}$. The power set of all possible states is $\text{P}(\Theta) = \{\{\text{fund}\}, \{\text{reject}\}, \Theta = \emptyset\}$.

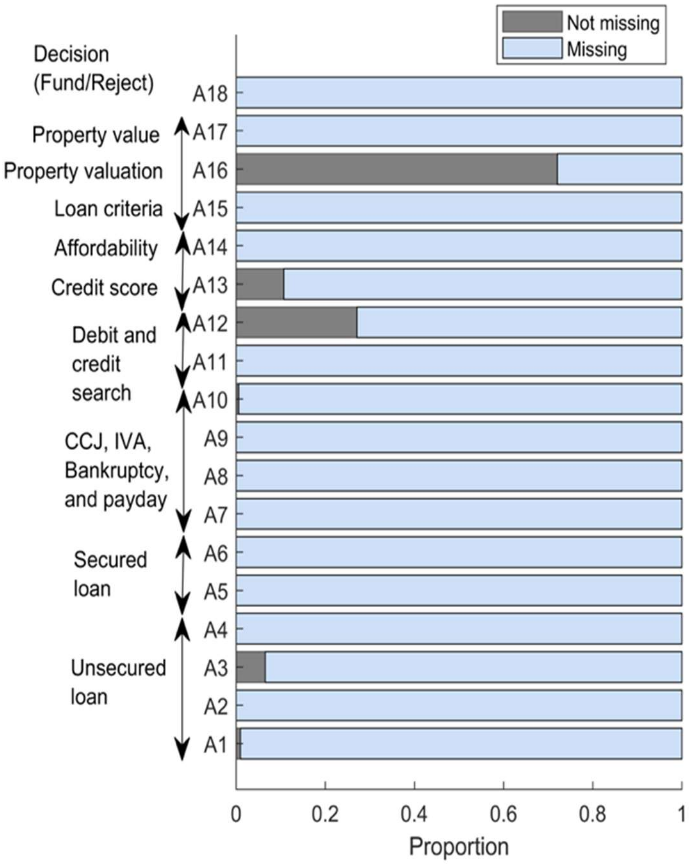


Figure 9. The proportion of missing values in mortgage loan dataset

The probability mass for all evidence in the data is estimated for all states in $P(\Theta)$ in order to implement I-MAKER and C-MAKER. The procedure to analyse the performance and sensitivity of I-MAKER and C-MAKER with other data imputation methods was similar to that of the asthma data. The data imputed by six different methods were analysed for three different missing data scenarios in three types of machine learning models. The hyper-parameters of ANN, decision tree, the BRB structure for individual attributes transformed by I-MAKER, and combined by C-MAKER is shown in Table 30, Table 32, Figure 17, and Figure 18, respectively. All attributes were fused into six groups by C-MAKER and I-MAKER transformed individual attributes without data fusion, can be seen in Figure 17 and 18, respectively. Table 28 in Appendix C shows the proportion of interrelated evidence, joint probability, and sparse index for each group for data fused by C-MAKER. Table 30 shows that the input data transformed for the ANN by C-MAKER has $17 \times 2 = 34$ and I-MAKER has $6 \times 2 = 12$ transformed features compared to 60 for other methods in scenario I. For scenarios II and III, the input data transformed for ANN by C-MAKER has $17 \times 3 = 54$ and I-MAKER has $6 \times 3 = 18$ transformed features compared to 60 for the other methods.

The AUC score of ANN, decision tree, and BRB for missing data imputed by six different methods under three missing data scenarios: 6.55%, 26.55% (6.55%+20%), and 41.55% (6.55%+35%) are shown in Table 15. The proportion of missingness in the data increased from scenario I to scenario II, which reduced the abstraction capacity of all the models. For all three scenarios, the performance of C-MAKER was best. Its performance was close to either missForest or I-MAKER in all scenarios. It is interesting to observe that in the asthma data, the performance of I-MAKER was better than other methods, especially C-MAKER, due to high sparsity in the join evidence when the pieces of evidence in two or more attributes were combined. However, in the mortgage loan data, C-MAKER is mostly better or very close to I-MAKER. One possible explanation for this outcome is that there are some attributes in the model that strongly influenced the prediction, so the loss of information of other attributes did not affect the result. In scenarios II and III, I-MAKER and C-MAKER have relatively similar performances. The relatively common order of AUC for scenario III, in which 45.99% of data was missing, is C-MAKER > I-MAKER > missForest > MICE > EM > KNN.

Table 15: AUC Performance comparison for mortgage loan data

| | **Scenario I**<br>**Original data missing proportion = 6.55%** | | | | | |
|---|---|---|---|---|---|---|
| | **I-MAKER** | **C- MAKER** | **missForest** | **MICE** | **EM** | **KNN** |
| **ANN** | 0.951 | 0.956 | 0.953 | 0.946 | 0.940 | 0.941 |
| **Decision tree** | 0.867 | 0.861 | 0.878 | 0.852 | 0.843 | 0.851 |
| **BRB** | 0.929 | 0.966 | 0.900 | 0.981 | 0.902 | 0.921 |
| **Scenario II**<br>**Missing proportional: original data% + 20% = 26.55%** | | | | | | |
| **ANN** | 0.856 | 0.858 | 0.794 | 0.780 | 0.712 | 0.701 |
| **Decision tree** | 0.780 | 0.790 | 0.730 | 0.757 | 0.751 | 0.752 |
| **BRB** | 0.863 | 0.869 | 0.741 | 0.711 | 0.7 | 0.853 |
| **Scenario III**<br>**Missing proportional: original data% + 35% = 41.55%** | | | | | | |
| **ANN** | 0.686 | 0.686 | 0.581 | 0.541 | 0.507 | 0.478 |
| **Decision tree** | 0.601 | 0.603 | 0.490 | 0.512 | 0.549 | 0.422 |
| **BRB** | 0.660 | 0.612 | 0.578 | 0.571 | 0.576 | 0.424 |

## 7.4 Discussion

### 7.4.1 Uncertainty Management

The results of both case studies demonstrate that I-MAKER, C-MAKER, and missForest has relatively close performance for three types of AI algorithms compared to other data imputation techniques. MissForest have been implemented widely and proven to be robust due to its ability to handle various types of data and less requirement of tuning (Waljee A. K., et al., 2013). There is no doubt that data imputation technique such as missForest based on random forest algorithm and any other ML algorithm for missing data imputation would provide a close cut performance. However, trust in transformed data is equally essential due to the rise in ethical concerns around data and computerized decisions by data-driven systems. It is found and argued that naturally meaningful features in structured data extracted by following a standard knowledge discovery process would not exhibit a significant difference in the performance of complex algorithms and simple algorithms after preprocessing (Rudin, 2019). This implies that features fed in the AI system should be relevant and trustworthy to strike a balance between model performance and overall explainability of an AI system. The role of transformed data in the explainability of AI models will be discussed in detail in the next section.

Missing values in input attributes and output attribute introduce uncertainty in the predicted decisions. The multiple pieces of evidence in an input attribute is mapped into output attribute space by MAKER rule. Similarly, multiple pieces of joint evidence in a combined attributes are mapped into output attribute space by conjunctive MAKER. Both techniques are based on DS theory. It is has an excellent ability to consider uncertainty and ignorance. It can cover uncertainty in decisions by local and global ignorance represented by subsets of singleton decisions and universal set of decisions, shown in Figure 10. Local ignorance refers to the cases where the evidence points to two or more decisions. In other words, they are partial states of the system. For example, experts would prefer to provide a subjective assessment for a subset of decisions than for one single decision (Xu, Yang, & Wang, 2006). Global ignorance refers to the cases where the evidence state (or possible outcome) is entirely unknown. In MAKER rule, the outcome is profiled over the subset of the power set, which represents the dimensions in output attribute space. The uncertainty in the evidence at an input attribute is reflected in uncertainty in the output profiled over the power set. In a practical system, evidence point to singleton decision and universal set of decisions (global ignorance). It is rare to find a dataset where an instance point to a subset of decisions (local ignorance). For example, if there are four stages of an illness. Then, symptoms in the dataset could point to early two stages or two later stages. Such dataset could exist when the data is labelled or annotated by the experts. In the case study, it is assumed that the evidence points to a singleton set of decisions and a universal set of decisions. Both methods can be used if the evidence point to subset of decisions. The uncertainty caused by importance of the evidence and sufficiency of the evidence is incorporated in MAKER through evidence weight and reliability, respectively. The weight of evidence can be trained or can be obtained by the subjective judgment by the domain experts.

$$2^{\theta} = P(\theta) = \{\emptyset, \overbrace{\{\theta_1\}, \dots, \{\theta_Z\}}^{\text{Singleton set}}, \underset{\text{Local ignorance}}{\overbrace{\{\theta_1, \theta_2\}, \dots, \{\theta_1, \theta_Z\} \dots, \{\theta_1, \dots, \theta_{Z-1}\}}^{\text{Subset}}}, \underset{\text{Global ignorance}}{\overbrace{\theta}^{\text{Universal set}}}\}$$

Figure 10. Local and global uncertainty

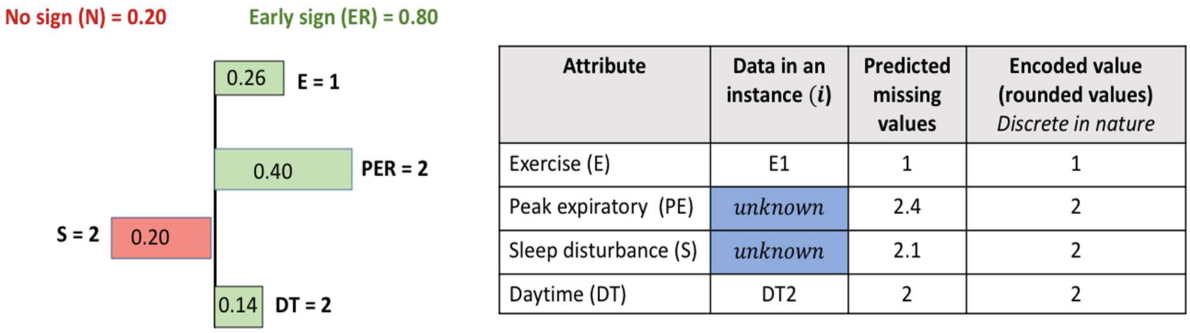
The EM data imputation techniques do not consider uncertainty (Baneshi & Talei, 2012), whereas MICE, KNN, and missForest can consider uncertainty by introducing a random component in the estimated values (Baneshi & Talei, 2011). These techniques cannot estimate the uncertainty of incomplete and ambiguous data, where both input and output attributes have missing values. The missing data is imputed in a cyclic fashion for each attribute in both MICE and missForest by predicting the missing values with only the complete instances in the dataset. In MAKER, the interrelation and basic probability are estimated with the complete case data; however, probability mass is estimated by both complete and incomplete data (entire dataset). It takes advantage of DS uncertainty principles.

### 7.4.2 Impact of Data Transformation on Explainability of Decisions

AI algorithms can parse a large amount of data into intelligent insights and predictions. These algorithms have efficient optimizers and have huge parametric space which results in complex black-box models which produce non-traceable decisions. Despite the unlimited potential of these algorithms, humans are still baffled how a black-box algorithm arrives at a particular decision. This question often raises concern about the reliability of autonomous decision-making systems. Explanation based on logical reasoning behind a decision is critical for domains like healthcare and finance where automated decisions have a high impact on human life. The explainability of AI/ML system has two stages: pre-modelling and post modelling. The pre-modelling explainability demands an in-depth understanding of data in a domain. The post-modelling explainability requires reasoning behind local (single decision) decisions and global understanding of the model. The simple approximation of decision boundaries of a black-box model by probing a trained AI/ML model with test dataset is a common strategy to understand a local decision by techniques like Local Interpretable Model Agnostic Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016) and Shapley (Lundberg & Lee, 2017). The explanation for local decisions and global understanding of the model can be approximated by model specific and model agnostic techniques (Adadi & Berrada, 2018). The pros and cons of these techniques can be seen in (Kelly L. , et al., 2020) (Adadi & Berrada, 2018) . The white-box models are inherently interpretable such as rule-based, decision tree, and linear regression.

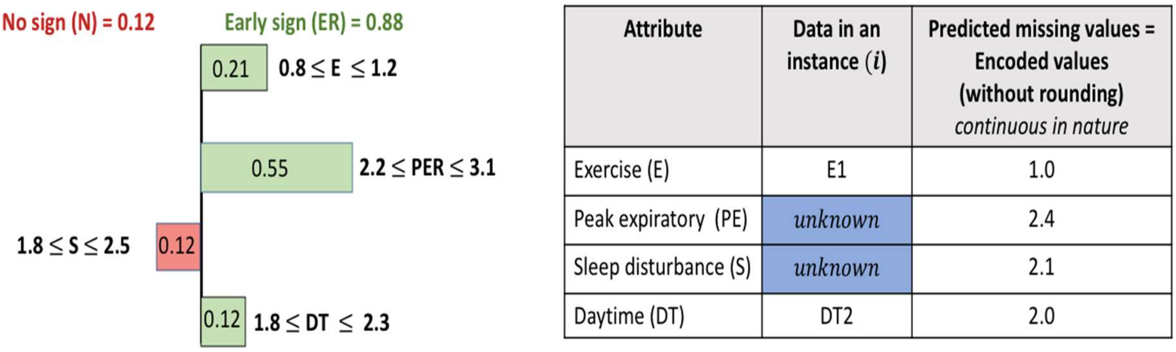
The algorithm follows the data. It learns biases and unknown uncertainty in from incomplete training data reflecting historical discrepancies, which could result in untrustworthy outcomes for a certain group of people. It could have an adverse impact on groups which are unrepresented in the training data without a developer's intention to discriminate. The group of people in the case study are asthma patients and loan customers. Therefore, it is essential to control and understand the uncertainty in incomplete and ambiguous data. This paper addresses the concerns in handling uncertainty in incomplete and ambiguous categorical attributes. Before predicting missing values, data imputation methods transform categorical attributes into numerical features containing discrete values ($\{1,2,3,\ldots\}$ or $\{0,1\}$). Similarly, AI/ML algorithm expects discrete values for a categorical attribute. However, the missing values predicted by data imputation methods are continuous. A discrete value points to a specific category in an attribute, whereas continuous value does not point to a specific category. A continuous value for a category can be perceived as a value which could belong to two categories with varying degree. For example, the categories in attribute $PER$ in asthma data encoded as $\{PER1 = 1, PER2 = 2, PER3 = 3\}$. A predicted value 2.4 points 60% towards a category $PER2$ and 40% towards category $PER3$. The continuous values are rounded to nearest discrete value to point specifically to a category for deep learning and tree-based models. This practice is helpful in post-modelling explainability stage. Also, for the transformation of data to a required format, for example, one-hot encoding for deep-learning models, and label encoding for tree-based models.

The reasoning behind the predicted values provides a clear understanding of the input data. It is an initial step to achieve the explainable system. Among all techniques, missForest has demonstrated relatively good performance compared to other methods (Stekhoven & Bühlmann, 2012) (Waljee A. , et al., 2013). It is based on the random-forest algorithm, which is not inherently interpretable; therefore, reasoning behind the imputation of missing values can only be explained by external methods. Similarly, KNN is not sufficiently interpretable as each cluster is defined by mean and covariance parameters. The distance from centroid or boundary of each cluster is usually used to determine the degree of membership for a given instance. However, it is not easy to understand the contribution of relevant features in the decision-making process, especially when the data has a large number of dimensions. The MICE imputation method implements an appropriate regression method for different types of attributes such as linear regression, logistic regression, multinomial log ‐ linear models, or Poisson regression (Gelman, Van Mechelen, Verbeke, Heitjan, & Meulders, 2005). A predicted missing value by the regression model could be explained by multiplying the weight of feature or coefficient with the dependent values (known/non-missing features). The Figure 11 and Figure 12 demonstrates the explainability of a probabilistic decision $\{(ER, 0.80), (N, 0.20)\}$ and $\{(ER, 0.88), (N, 0.12)\}$ given by ANN for an instance by feature importance obtained from model-agnostic method LIME, respectively. The missing data in both the cases are predicted by missForest, therefore the pre-modelling explainability would be very explicit. It is hard to understand how and why these values are predicted. The importance of explainability is described simply on asthma data for due to small attributes compared to mortgage loan.



| Attribute | Data in an instance ($i$) | Predicted missing values | Encoded value (rounded values) *Discrete in nature* |
|---|---|---|---|
| Exercise (E) | E1 | 1 | 1 |
| Peak expiratory (PE) | *unknown* | 2.4 | 2 |
| Sleep disturbance (S) | *unknown* | 2.1 | 2 |
| Daytime (DT) | DT2 | 2 | 2 |

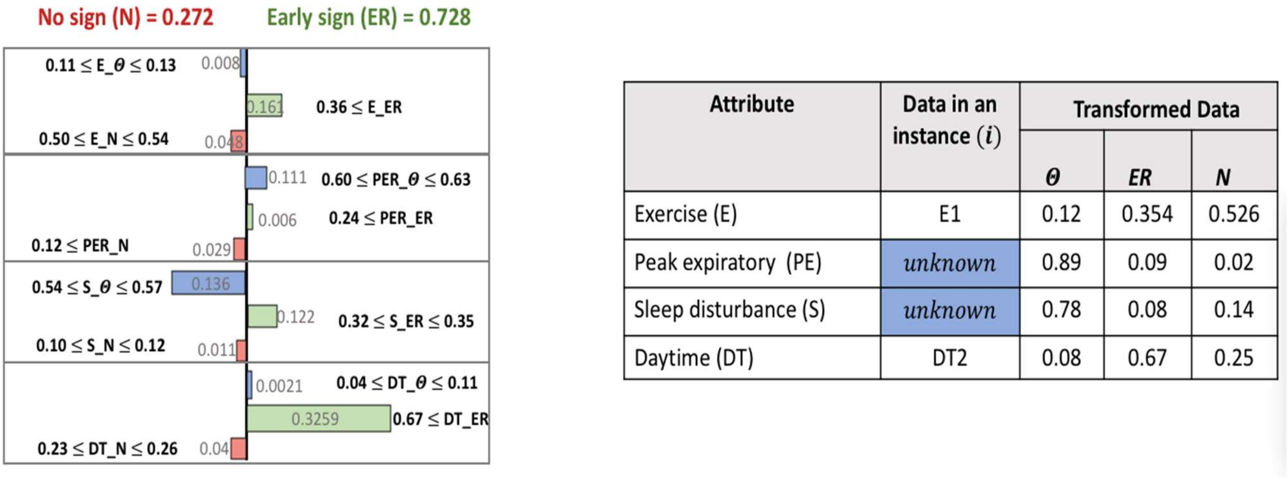
Figure 11. Explainability by feature importance for an instance when imputed data is rounded



| Attribute | Data in an instance ($i$) | Predicted missing values = Encoded values (without rounding) *continuous in nature* |
|---|---|---|
| Exercise (E) | E1 | 1.0 |
| Peak expiratory (PE) | *unknown* | 2.4 |
| Sleep disturbance (S) | *unknown* | 2.1 |
| Daytime (DT) | DT2 | 2.0 |

Figure 12. Explainability by feature importance for an instance when imputed data is not rounded

In Figure 11, the predicted missing values were approximated (or rounded) to create discrete input data to point explicitly to the categories in the data. It provides better support in the post-modelling explainability stage. In Figure 12, the predicted missing values in the input data (all four attributes in asthma data) were not approximated to a nearest discrete value. The input data in such case cannot be decoded back to the original categories, and model-agnostic method assumes that all four variables are continuous. It is difficult to debate about the input data requirements for post-modelling explainability. The discrete data can point specifically to a category, thus can provide explainability in terms of contribution by each attribute towards a decision which point to a specific category, whereas continuous data is real predicted value and can provide contributions towards a decision by the attributes which falls in a certain range for each instance. Figure 13 demonstrate the explainability of a probabilistic decision $\{(ER, 0.728), (N, 0.272)\}$ given by ANN for an instance by feature importance obtained from model-agnostic method LIME. The table in shown in Figure 13 is the input data transformed by MAKER. Here, a decision for an instance can be broken down into the contribution of each feature towards $\{ER, N, \Theta = \{ER, N\}\}$, where $\Theta$ represent uncertainty. The simplified explanation for the decision for I-MAKER data is shown in Figure 14. The visualization of the importance of the features by C-MAKER would be same as I-MAKER.



| Attribute | Data in an instance ($i$) | Transformed Data | | |
|---|---|---|---|---|
| | | $\theta$ | ER | N |
| Exercise (E) | E1 | 0.12 | 0.354 | 0.526 |
| Peak expiratory (PE) | *unknown* | 0.89 | 0.09 | 0.02 |
| Sleep disturbance (S) | *unknown* | 0.78 | 0.08 | 0.14 |
| Daytime (DT) | DT2 | 0.08 | 0.67 | 0.25 |

Figure 13. Explainability by feature importance for an instance when data is imputed by I-MAKER

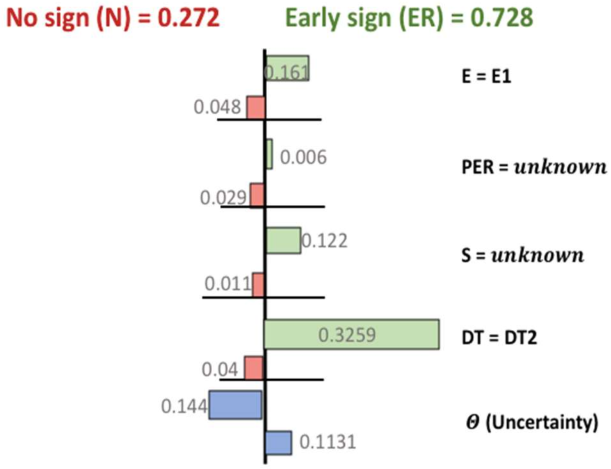


Figure 14. Simplified explanation by feature importance for an instance when data is imputed by I-MAKER

## 8 Conclusion

Automated decision-making systems cannot anticipate every circumstance due to uncertainty induced by incomplete and ambiguous training data reflecting historical discrepancies and imperfections in the data collection process. A meaningful trusted relationship between humans and decision-making systems is critical for successful adoption of these systems to provide automated decisions in highly regulated domains. Missing data has a ubiquitous presence in a realistic setting. The first step to engender human trust demands consideration of uncertainty due to missing information and a firm understanding of input data pre-processed by imputation and transformation techniques. An inadequate understanding of input data by humans prevents the achievement of a safe and trustworthy decision-making process.

This paper comprehensively addresses the issue of uncertainty in decision-making due to categorical attributes. The categorical attributes have inherently non-numerical nature compared to numerical or continuous attributes. The inherent non-numerical nature and presence of incomplete and ambiguous values in categorical attributes increase the uncertainty in decision-making. This paper has recognised three sources of uncertainties in categorical attributes. The informational uncertainty, unforeseeable uncertainty in the decision task environment, and the uncertainty due to lack of pre-modelling explainability in categorical attributes are addressed in the proposed methodology in this paper.

The MAKER rule is proposed for an interpretable numerical transformation and imputation of incomplete and ambiguous categorical attributes. It integrates the recognised uncertainties in the transformed input data that allow a model to perceive data limitations and acknowledge doubtful predictions during the training regime. MAKER can be implemented on individual categorical attributes and can combine two or more attributes to reduce the dimensionality of transformed numerical features. It does not discard instances with missing values for complete-case analysis like most data imputation techniques. It imputes missing values and numerically transforms categorical attributes by analysing all complete and incomplete instances to incorporate uncertainty. This paper has demonstrated that the uncertainty management and interpretable transformation of categorical attributes by MAKER provide support for trustworthy pre-modelling and post-modelling explainability to understand the input data and reasoning behind a decision, respectively. Additionally, it has a notion of weight and reliability of evidence for each outcome to include the subjective preference of an expert over a piece of evidence and the quality of the evidence in a categorical attribute, respectively. MAKER rule in this paper has not addressed the approach to combine subjective judgment of multiple human experts at different levels of expertise in a domain. It expects an expert to provide crisp numerical judgment on the weight of the evidence between [0,1]. However, the subjective judgment from an expert may not be reasonable. Another disadvantage of MAKER is that it can either utilize evidence weight as a subjective judgment from an expert or be treated as a parameter trained by data-driven optimization. The weight of evidence is trained if subjective judgments are not available. It does not provide a reasonable proposition to choose the weight of evidence if both are subjective judgments and the ability to train is available. It neither provides the ability to combine both sources of information. In the future, MAKER can be extended to address the ambiguity in judgment by multiple human experts. Its scope can be broadened by expanding it to pre-process both categorical and continuous data.

The practicality of the proposed methodology is demonstrated on paediatric asthma symptoms data collected from National Health Services in the UK and mortgage loan data obtained from a lending firm in the UK. The MAKER for an individual attribute (I-MAKER) and for the fusion of a group of attributes (C-MAKER) were compared with four widely implemented data imputation techniques: missForest, MICE, EM, and KNN. The performance and sensitivity of these methods were analysed on asthma and loan data for three different missing data scenarios by three types of AI models- artificial neural network (deep-learning), decision-tree (tree-based), and belief-rule-base (rule-based). The experimental results demonstrated that the proposed methods outperformed other data imputation techniques in most scenarios for different AI models. Among the various scenarios and AI models, C-MAKER achieved the highest AUC value in the mortgage loan data and I-MAKER in the paediatric asthma data.

This research would enable developers to design AI-enabled decision-making systems that can integrate uncertainty in decision-making and support the explainability of decisions by black-box and white-box models. The understanding of the limitations of data by uncertainty management allows the generation of smarter and reliable computerised decisions for highly regulated domains in high-risk situations.

**Acknowledgements:**

We are thankful to National Health Services (NHS), UK for collection and dissemination of asthma signs & symptoms data. We are thankful to Together Financial Services for providing the mortgage loan data. This research work is the output of DART project funded by Alliance Strategic Research Investment Grant (Grant Code: AA15032) at The University of Manchester, UK.

**References:**

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138-52160.

Agrawal, A., Gans, J., & Goldfarb, A. (2020). How to win with machine learning : and how to catch up if you're lagging behind. *Harvard Business Review*, 126-133.

Alkharusi, H. (2012). Categorical variables in regression analysis: A comparison of dummy and effect coding. *International Journal of Education*, 202–210.

Almaghrabi, F., Xu, D. L., & Yang, J. B. (2019). A new machine learning technique for predicting traumatic injuries outcomes based on the vital signs. *25th International Conference on Automation and Computing (ICAC)* (pp. 1-5). IEEE.

Almaghrabi, F., Xu, D.-L., & Yang, J.-B. (2021). An evidential reasoning rule based feature selection for improving trauma outcome prediction. *Applied Soft Computing*, 107112.

Audigier, V., Husson, F., & Josse, J. (2017). MIMCA: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and computing*, 501-518.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Azur, Melissa J., Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. "Multiple imputation by chained equations: what is it and how does it work? 40-49.

Baneshi, M. R., & Talei, A. R. (2011). Multiple imputation in survival models: applied on breast cancer data. *Iranian Red Crescent Medical Journal*, 544.

Baneshi, M., & Talei, A. (2012). Does the missing data imputation method affect the composition and performance of prognostic models? *Iranian Red Crescent Medical Journal*, 31.

Bogosian, K. (2017). Implementation of moral uncertainty in intelligent machines. *Minds and Machines*, 591-608.

Bourgeois III, L. J. (1980). Strategy and environment: A conceptual integration. *Academy of management review*, 25-39.

Briggs, A., Clark, T., Wolstenholme, J., & Clarke, P. (2003). Missing.... presumed at random: cost-analysis of incomplete data. *Health economics*, 377-392.

Brown, R. L. (1994). Brown, Roger L. "Efficacy of the indirect approach for estimating structural equation models with missing data: A comparison of five methods. *Structural Equation Modeling: A Multidisciplinary Journal* , 287-316.

Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 1477-1494.

Cheng, C. Y., Tseng, W. L., Chang, C. F., Chang, C. H., & Gau, S. S. (2020). A Deep Learning Approach for Missing Data Imputation of Rating Scales Assessing Attention-Deficit Hyperactivity Disorder. *Frontiers in psychiatry*, 673.

Dempster, A. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 205-232.

Dempster, A. (2008). *Upper and lower probabilities induced by a multivalued mapping.* Berlin, Heidelberg: Springer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1-22.

Deng, Y., Chang, C., Ido, M. S., & Long, Q. (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Scientific reports*, 1-10.

Dubois, D., & Prade, H. (1988). Representation and combination of uncertainty with belief functions and possibility measures. *Computational intelligence*, 244-264.

Ducange, P., Pecori, R., & Mezzina, P. (2018). A glimpse on big data analytics in the framework of marketing strategies. *Soft Computing*, 325-342.

Fisher, R. A. (1992). *Statistical methods for research workers.* New York, NY: Springer.

Fleiss, J., Levin, B., & Paik, M. (2013). *Statistical methods for rates and proportions.* john wiley & sons.

Frikha, A., & Moalla, H. (2015). Analytic hierarchy process for multi-sensor data fusion based on belief function theory. *European Journal of Operational Research , 241*(1), 133-147.

Ganji, S. S., Abbas Rassafi, A., & Jamshidi Bandari, S. (2020). Application of evidential reasoning approach and OWA operator weights in road safety evaluation considering the best and worst practice frontiers. *Socio-Economic Planning Sciences*, 100706.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D., & Meulders, M. (2005). Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics*, 74-85.

Goodman, L. A. (1971). Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables. *Journal of the American statistical Association*, 339-344.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, 197-218.

Han, Y., Jiao, J., & Weissman, T. (2015). Does Dirichlet prior smoothing solve the Shannon entropy estimation problem? *IEEE International Symposium on Information Theory (ISIT)* (pp. 1367-1371). IEEE.

Hughes, R., Heron, J., Sterne, J., & Tilling, K. (2019). Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International journal of epidemiology*, 1294-1304.

Huque, M., Carlin, J., Simpson, J., & Lee, K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology volume, 18*(168).

Iselin, E. (1989). The impact of information diversity on information overload effects in unstructured managerial decision making. *Journal of Information Science*, 163-173.

Jamshidian, M., & Mata, M. (2007). Advances in analysis of mean and covariance structure when data are incomplete. In *Handbook of latent variable and related models* (pp. 21-44). North-Holland.

Kelly, L., Sachan, S., Ni, L., Almaghrabi, F., Allmendinger, R., & Chen, Y. (2020). Explainable Artificial Intelligence for Digital Forensics: Opportunities, Challenges and a Drug Testing Case Study. In *Digital Forensic Science.* IntechOpen.

Kelly, L., Sachan, S., Ni, L., Almaghrabi, F., Allmendinger, R., & Chen, Y. W. (2020). Explainable Artificial Intelligence for Digital Forensics: Opportunities, Challenges and a Drug Testing Case Study. In *Digital Forensic Science.* IntechOpen.

Lan, Q., Xu, X., Ma, H., & Li, G. (2020). Multivariable data imputation for the analysis of incomplete credit data. *Expert Systems with Applications , 141.*

Langan, R., Archibald, R., & Lamberti, V. (2016). Nuclear forensics analysis with missing data. *Journal of Radioanalytical and Nuclear Chemistry*, 687-692.

Lipshitz, R., & Strauss, O. (1997). Coping with uncertainty: A naturalistic decision-making analysis. *Organizational behavior and human decision processes*, 149-163.

Liu, X., Sachan, S., Yang, J.-B., & Xu, D.-L. (2019). Maximum Likelihood Evidential Reasoning-Based Hierarchical Inference with Incomplete Data. *In 2019 25th International Conference on Automation and Computing (ICAC)* (pp. 1-6). IEEE.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

Maddulapallia, A. K., Yang, J.-B., & Xu, D.-L. (2012). Estimation, modeling, and aggregation of missing survey data for prioritizing customer voices. *European Journal of Operational Research, 220*(3), 762-776.

Masconi, K., Matsha, T., Echouffo-Tcheugui, J., Erasmus, R. T., & Kengne, A. P. (2015). Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus: a systematic review. *EPMA Journal, 6*(7).

Meng, X. L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 899-909.

Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter 3*, (pp. 27-32).

Milliken, F. J. (1987). Three types of perceived uncertainty about the environment: State, effect, and response uncertainty. *Academy of Management review*, 133-143.

Qin, B., Xia, Y., & Prabhakar, S. (2011). Rule induction for uncertain data. *Knowledge and information systems*, 103-130.

Razavi-Far, R., Chakrabarti, S., Saif, M., & Zio, E. (2019). An integrated imputation-prediction scheme for prognostics of battery data with missing observations. *Expert Systems with Applications*, 709-723.

Redshaw, C. H., Stahl-Timmins, W. M., & Fleming, L. E. (2013). Potential changes in disease patterns and pharmaceutical use in response to climate change. *Journal of Toxicology and Environmental Health, Part B*, 285-320.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), arXiv preprint arXiv:1606.05386.* New-York.

Richman, M. B., Trafalis, T. B., & Adrianto, I. (2009). Missing data imputation through machine learning algorithms. In *Artificial intelligence methods in the environmental sciences* (pp. 153-169). Dordrecht: Springer.

Roth, P. L. (1994). MISSING DATA: A CONCEPTUAL REVIEW FOR APPLIED PSYCHOLOGISTS. *Personnel Psychology, 47*, 537-560.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 581-592.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence , 1*(5), 206-215.

Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. *Journal of machine learning research*, 1623-1657.

Sachan, S., Yang, J. B., & Xu, D. L. (2020). Global and local interpretability of belief rule base. *In Developments Of Artificial Intelligence Technologies In Computation And Robotics-Proceedings Of The 14th International Flins Conference* (p. p. 68.). Hamburg: World Scientific.

Sachan, S., Yang, J., Xu, D., Benavides, D., & Li, Y. (2020). An explainable AI decision-support-system to automate loan underwriting. *Expert Systems with Applications*, 113100.

Sachan, S., Zhou, C., Wen, R., Sun, W., & Song, C. (2016). Multiple correspondence analysis to study failures in a diverse population of a cable. *IEEE Transactions on Power Delivery*, 1696-1704.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* CRC press.

Schelter, S., Biessmann, F., Januschowski, T., Salinas, D., Seufert, S., Szarvas, G., & Deshpande, A. (2020). Challenges in Machine Learning Model Management. *IEEE Data Eng. Bull.*, 5-15.

Shafer, G. (1976). *A mathematical theory of evidence* (Vol. 42). Princeton university press.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American journal of epidemiology*, 764-774.

Sidi, Y., & Harel, O. (2018). The treatment of incomplete data: reporting, analysis, reproducibility, and replicability. *Social Science & Medicine*, 169-173.

Sim, J., Kwon, O., & Lee, K. C. (2016). Adaptive pairing of classifier and imputation methods based on the characteristics of missing values in data sets. *Expert Systems with Applications*, 485-493.

Simonoff, J. S. (1995). Smoothing categorical data. *Journal of Statistical Planning and Inference, 47*(1-2), 41-69.

Smarandache, F., Dezert, J., & Tacnet, J. M. (2010). Fusion of sources of evidence with different importances and reliabilities. *13th International Conference on Information Fusion* (pp. 1-8). IEEE.

Smets, P., & Kennes, R. (1994). The transferable belief model. *Artificial intelligence*, 191-234.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 112-118.

Tang, S. W., Zhou, Z. J., Hu, C. H., Zhao, F. J., & Cao, Y. (2020). A New Evidential Reasoning Rule-Based Safety Assessment Method With Sensor Reliability for Complex Systems. *IEEE Transactions on Cybernetics*.

Tang, Y., Zhou, D., He, Z., & Xu, S. (2017). An improved belief entropy–based uncertainty management approach for sensor data fusion. *International Journal of Distributed Sensor Networks, 13*(7).

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

Von Eye, A., & Clogg, C. e. (1996). Categorical variables in developmental research. *Methods of analysis*.

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., & Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*.

Waljee, A., Mukherjee, A., Singal, A., Zhang, Y., Warren, J., Balis, U., . . . Higgins, P. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*.

Walker, W., Haasnoot, M., & Kwakkel, J. (2013). Adapt or perish: a review of planning approaches for adaptation under deep uncertainty. *Sustainability*, 955-979.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 86-95.

Wang, G., Zhang, X., Wang, H., Chu, Y., & Shao, Z. (2021). Group-Oriented Paper Recommendation With Probabilistic Matrix Factorization and Evidential Reasoning in Scientific Social Network. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

Wang, H., & Wang, S. (2009). Discovering patterns of missing data in survey databases: An application of rough sets. *Expert Systems with Applications, 36*(3-2), 6256-6260.

Weinberger, K., Dasgupta, A., Langford, J., Smola, A., & Attenberg, J. (2009). Feature hashing for large scale multitask learning. *Proceedings of the 26th annual international conference on machine learning*, (pp. 1113-1120).

Wu, J., & Shang, S. (2020). Managing Uncertainty in AI-Enabled Decision Making and Achieving Sustainability. *Sustainability*, 8758.

Xu, D.-L., Yang, J.-B., & Wang, Y.-M. (2006). The evidential reasoning approach for multi-attribute decision analysis under interval uncertainty. *European Journal of Operational Research, 174*(3), 1914-1943.

Xu, X., Zheng, J., Yang, J.-b., Xu, D.-l., & Chen, Y.-w. (2017). Data classification using evidence reasoning rule. *Knowledge-Based Systems*(116), 144-151.

Yager, R. R. (1987). On the Dempster-Shafer framework and new combination rules. *Information sciences*, 93-137.

Yang, J. B. (2001). Rule and utility based evidential reasoning approach for multiattribute decision analysis under uncertainties. *European journal of operational research, 131*(1), 31-61.

Yang, J. B., & Xu, D. L. (2017). Inferential modelling and decision making with data. *23rd International Conference on Automation and Computing (ICAC)* (pp. 1-6). IEEE.

Yang, J. B., Liu, J., Wang, J., Sii, H. S., & Wang, H. W. (2006). Belief rule-base inference methodology using the evidential reasoning approach-RIMER. *IEEE Transactions on systems, Man, and Cybernetics-part A: Systems and Humans*, 266-285.

Yang, J.-B., & Xu, D.-L. (2002). On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 32*(3), 289-304.

Yang, J.-B., & Xu, D.-L. (2013). Evidential reasoning rule for evidence combination. *Artificial Intelligence*(205), 1-29.

Yang, L. H., Wang, S., Ye, F. F., Liu, J., Wang, Y. M., & Hu, H. (2021). Environmental investment prediction using extended belief rule-based system and evidential reasoning rule. *Journal of Cleaner Production*, 125661.

Zhang, M.-J., Wang, Y.-M., Li, L.-H., & Chen, S.-Q. (2017). A general evidential reasoning algorithm for multi-attribute decision analysis under interval uncertainty. *European Journal of Operational Research* , *257*(3), 1005-1015.

## Appendix A. Algorithmic Steps

Table 16: MAKER Algorithm to Pre-Process Individual Categorical Attributes

**I-MAKER Algorithm: MAKER to Pre-Process Individual Categorical Attributes**

```
Random stratified split of a dataset $\mathcal{D}$ into cross validation $\mathcal{D}^{CV}$ and validation set $\mathcal{D}^{vld}$
Split $\mathcal{D}^{CV}$ into $K$ equal folds
Create $K$ different training set $\mathcal{D}^{train,k}$ and test set $\mathcal{D}^{test,k}$ from $K$ folds
$P_{store} \longleftarrow \phi$ //Initialize empty list of probability mass//
For $q = 1$ to $Q$ do:
    $P_{temp} \longleftarrow \phi$ //Initialize temporary list for probability mass//
    For $k = 0$ to $K$ do: //Each $k^{th}$ training and test set//
        Create Contingency table from $\mathcal{D}^{train,k}$
        Compute Likelihood
        Compute basic probability //Equation (5)//
        Compute prior probability //Equation (10)//
        Compute reliability of evidence pointing to class $P(\Theta)$ //Equation (7)//
        Assume initial weight = reliability
        Compute initial probability mass //Equation (6)//
        Update probability mass of $Unknown = \left\{A_1, \dots A_v, \dots A_{V_q}\right\}$ //Equation (11)//
          a) Test classification accuracy of $Unknown$ and find $\lambda_{\theta,q}$ for $\mathcal{D}^{test,k}$
          b) Adjust probability mass of $Unknown$ pointing to class $\theta$
        Optimize weight of evidence by finding $\widehat{m}(w)$ by iterating overstep (8) and (10) //This step can be avoided to save pre-training time//
        $P_{temp} \longleftarrow [\widehat{m}^{k}_{\theta,v,q}]$ //Store probability mass of all evidence in $q^{th}$ attribute obtained from $k^{th}$ set//
    End For
    $P_{store} \longleftarrow [\widehat{m}_{\theta,v,q}]$ //Store average probability mass//
End For
Numerical transformation of all $q^{th}$ ($q \in \{1, \dots, Q\}$) categorical attributes in $\mathcal{D}^{CV}$ and $\mathcal{D}^{vld}$
```

**Time complexity** = $O(Q) \times O\big(K \times O(training)\big)$

Table 17: Conjunctive MAKER Algorithm to Pre-Process and Combine Multiple Categorical Attributes

| | **C-MAKER Algorithm: Conjunctive MAKER to Pre-Process and Combine Multiple Categorical Attributes**<br>C-MAKER can combine into a maximum $\boldsymbol{G}$ number of categorical attributes in a dataset containing a total $\boldsymbol{Q}$ number of categorical attributes, such that $\boldsymbol{G} < \boldsymbol{Q}$. The $\boldsymbol{G}$ number of attributes can be combined in $\boldsymbol{\Omega}$ number of ways (Equation (12)). A set of the most feasible combinations with high interrelation and evidence sparse index is $\overline{\overline{\boldsymbol{\Omega}}} = \{\{\boldsymbol{A}^{q'}, \boldsymbol{A}^{q'+1}\}, \dots, \{\boldsymbol{A}^{1}, \dots, \boldsymbol{A}^{q}, \dots, \boldsymbol{A}^{Q'}\}\}$; here cardinality of set $\overline{\overline{\boldsymbol{\Omega}}}$ is less than $\boldsymbol{\Omega}$. |
|---|---|
| 1 | Random stratified split of a dataset $\mathcal{D}$ into cross validation $\mathcal{D}^{CV}$ and validation set $\mathcal{D}^{vld}$ |
| 2 | Split $\mathcal{D}^{CV}$ into $K$ equal folds |
| 3 | Create $K$ different training set $\mathcal{D}^{train,k}$ and test set $\mathcal{D}^{test,k}$ from $K$ folds |
| 4 | Suppose, C-MAKER process a combination $\{\boldsymbol{A}^{1}, \dots, \boldsymbol{A}^{q}, \dots, \boldsymbol{A}^{Q'}\}$ out of $\Omega$ number of combinations in the following loop.<br>**For** $\omega = 1$ *to cardinality*$(\overline{\overline{\boldsymbol{\Omega}}})$ **do:** *//Loop process combinations feasible in set* $\overline{\overline{\boldsymbol{\Omega}}}$ *//* |
| 5 | Create a set of sets of singleton and sequential cumulative set of $\boldsymbol{Q}'$ attributes<br>$\mathbb{G} = \left\{\{\boldsymbol{A}^{1}\}, \dots, \{\boldsymbol{A}^{q}\}, \dots, \{\boldsymbol{A}^{Q'}\}, \dots, \{\boldsymbol{A}^{1}, \boldsymbol{A}^{2}\}, \dots, \{\boldsymbol{A}^{1}, \boldsymbol{A}^{q}\}, \dots, \{\boldsymbol{A}^{1}, \boldsymbol{A}^{2}, \boldsymbol{A}^{3}\}\}, \dots, \{\boldsymbol{A}^{1}, \dots, \boldsymbol{A}^{q}, \dots, \boldsymbol{A}^{Q'}\}\right\}$ |
| 6 | $P_{store} \longleftarrow \phi$ *//Initialize empty list of probability mass//* |
| 7 | **For** $k = 1$ to $K$ **do:** |
| 8 | $P_{temp} \longleftarrow \phi$ *//Initialize temporary list for probability mass//* |
| 9 | Extract complete data samples ($\mathcal{D}_{CM}$) for all set of attributes in $\mathbb{G}$ |
| 10 | *//Find missingness in the data //*<br>**If** $\mathcal{D}_{CM} = \mathcal{D}^{train,k}$ **then:** *//if complete data for a set of attributes in* $\mathbb{G}$ *is equal data in* $\mathcal{D}^{train,k}$ *//*<br>No missing data<br>**Else:**<br>missing data |
| 11 | Create Contingency table *//Created from* $\mathcal{D}_{CM}$ *data//* |
| 12 | Compute Likelihood |
| 13 | Compute basic probability *//Equation (5)//* |
| 14 | Compute Interrelation *//Equation (14a)//* |
| 15 | Compute reliability of evidence pointing to class $P(\Theta)$ *//Equation (7)//* |
| 16 | Compute reliability of evidence *//Equation (18)//* |
| 17 | Assume initial weight = reliability |
| 18 | Compute initial probability mass for singleton attributes in $\mathbb{G}$ from $\mathcal{D}^{train,k}$ data *//Equation (6)//* |
| 20 | Calculate combine probability mass *//Equation (17a)-(17c)//* |
| 21 | Optimize weight and reliability ratio by iterating overstep (17) |
| 22 | $P_{temp} \longleftarrow [\widehat{m}^{k}_{\theta,v1,\dots,VQ'}]$ *//Store probability mass of all evidence in* $q^{th}$ *attribute obtained from* $k^{th}$ *set //* |
| 23 | **End For** |
| 24 | $P_{store} \longleftarrow [\widehat{m}^{k}_{\theta,v1,\dots,VQ'}]$ *//Store average probability mass//* |
| 25 | **End For** |

Table 17: Conjunctive MAKER Algorithm to Pre-Process and Combine Multiple Categorical Attributes

| | |
|---|---|
| **C-MAKER Algorithm: Conjunctive MAKER to Pre-Process and Combine Multiple Categorical Attributes**<br>C-MAKER can combine into a maximum $\boldsymbol{G}$ number of categorical attributes in a dataset containing a total $\boldsymbol{Q}$ number of categorical attributes, such that $\boldsymbol{G} < \boldsymbol{Q}$. The $\boldsymbol{G}$ number of attributes can be combined in $\boldsymbol{\Omega}$ number of ways (Equation (12)). A set of the most feasible combinations with high interrelation and evidence sparse index is $\overline{\overline{\boldsymbol{\Omega}}} = \{\{\boldsymbol{A}^{q'}, \boldsymbol{A}^{q'+1}\}, \dots, \{\boldsymbol{A}^{1}, \dots, \boldsymbol{A}^{q}, \dots, \boldsymbol{A}^{Q'}\}\}$**;** here cardinality of set $\overline{\overline{\boldsymbol{\Omega}}}$ is less than $\boldsymbol{\Omega}$. | |
| 26 | Numerically transform and fuse all $\omega = 1$ to $cardinality(\overline{\overline{\boldsymbol{\Omega}}})$ number of combinations of categorical attributes into $Z$ or $Z + 1 - dimensional$ numerical features. |
| **Time complexity** = $O\big(cardinality\ of\ \overline{\overline{\Omega}}\big) \times O\big(K \times O(training)\big)$ | |

## Appendix B. Numerical Example on Real Data

Table 18: Likelihood of E and PE

| ***Physical exercise (E)*** | E1 | E2 | E3 |
|---|---|---|---|
| ***Early Diagnosis*** | | | |
| *Θ* | 0.882 | 0.059 | 0.059 |
| ER | 0.584 | 0.406 | 0.009 |
| N | 0.538 | 0.447 | 0.015 |

| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 |
|---|---|---|---|
| ***Early Diagnosis*** | | | |
| *Θ* | 0.412 | 0.529 | 0.059 |
| ER | 0.835 | 0.142 | 0.023 |
| N | 0.009 | 0.989 | 0.003 |

Table 19: Likelihood of joint evidences in E and PE

| ***Physical exercise (E)*** | E1 | | | E2 | | | E3 | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 |
| ***Early Diagnosis*** | | | | | | | | | |
| *Θ* | 0.308 | 0.538 | 0.077 | 0.000 | 0.000 | 0.000 | 0.000 | 0.077 | 0.000 |
| ER | 0.472 | 0.080 | 0.017 | 0.345 | 0.068 | 0.005 | 0.009 | 0.002 | 0.002 |
| N | 0.008 | 0.530 | 0.002 | 0.001 | 0.444 | 0.001 | 0.000 | 0.015 | 0.000 |

Table 20: Reliability of evidences in E and PE (Initial weight of evidences in E and PE)

| ***Physical exercise (E)*** | E1 | E2 | E3 |
|---|---|---|---|
| ***Early Diagnosis*** | | | |
| *Θ* | 0.01 | 0.00 | 0.03 |
| ER | 0.44 | 0.37 | 0.24 |
| N | 1.00 | 1.00 | 1.00 |

| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 |
|---|---|---|---|
| ***Early Diagnosis*** | | | |
| *Θ* | 0.01 | 0.00 | 0.06 |
| ER | 1.00 | 0.04 | 1.00 |
| N | 0.03 | 1.00 | 0.38 |

Table 21: Reliability of joint evidences in E and PE (Initial weight of joint evidences in E and PE)

| ***Physical exercise (E)*** | E1 | | | E2 | | | E3 | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 |
| ***Early Diagnosis*** | | | | | | | | | |
| *Θ* | 0.014 | 0.006 | 0.100 | 0.000 | 0.000 | 0.000 | 0. 000 | 0.032 | 0.000 |
| ER | 1.000 | 0.042 | 1.000 | 1.000 | 0.043 | 1.000 | 1.000 | 0.032 | 1.000 |
| N | 0.058 | 1.000 | 0.400 | 0.015 | 1.000 | 0.667 | 0.000 | 1.000 | 0.000 |

Table 22: Initial reliability ratio of joint evidences in E and PE

| ***Physical exercise (E)*** | E1 | | | E2 | | | E3 | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 |
| ***Early Diagnosis*** | | | | | | | | | |
| *Θ* | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ER | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| N | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Table 23: Trained weight of evidences in E and PE

| ***Physical exercise (E)*** | E1 | E2 | E3 |
|---|---|---|---|
| ***Early Diagnosis*** | | | |
| *Θ* | 0.01 | 0 | 0.03 |
| ER | 0.44 | 0.56 | 0.56 |
| N | 0.9 | 0.87 | 1 |

| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 |
|---|---|---|---|
| ***Early Diagnosis*** | | | |
| *Θ* | 0.01 | 0 | 0.09 |
| ER | 0.98 | 0.04 | 0.88 |
| N | 0.03 | 0.89 | 0.67 |

Table 24: Trained weight of joint evidences in E and PE

| ***Physical exercise (E)*** | E1 | | | E2 | | | E3 | | |
|---|---|---|---|---|---|---|---|---|---|
| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 |
| ***Early Diagnosis*** | | | | | | | | | |
| *Θ* | 0.90 | 0.70 | 0.91 | 1.00 | 0.90 | 0.91 | 0.90 | 0.98 | 0.90 |
| ER | 0.98 | 0.36 | 0.90 | 1.00 | 0.18 | 1.00 | 0.90 | 0.88 | 0.98 |
| N | 0.59 | 0.97 | 0.93 | 0.79 | 0.98 | 0.98 | 1.00 | 1.00 | 0.97 |

Table 25: Trained reliability ratio of joint evidences in E and PE

| ***Physical exercise (E)*** | E1 | E2 | E3 |
|---|---|---|---|

| ***Peak expiration (PE)*** | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 | PER1 | PER2 | PER3 |
|---|---|---|---|---|---|---|---|---|---|
| ***Early Diagnosis*** | | | | | | | | | |
| *Θ* | 0.90 | 1 | 0.98 | 0.99 | 0.99 | 1 | 1 | 1 | 1 |
| ER | 1 | 1 | 0.96 | 1 | 0.99 | 0.98 | 0.97 | 0.98 | 0.99 |
| N | 0.80 | 0.99 | 1 | 1 | 1 | 1 | 1 | 1 | 0.99 |

Table 26: Example of data fusion and transformation by MAKER

| | | | | | | ***Transformed data E&PE*** | | | ***Transformed data S&DT*** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # data points | ***E*** | ***PE*** | ***S*** | ***DT*** | | ***MK1_ Θ*** | ***MK1_ER*** | ***MK1_N*** | ***MK2_ Θ*** | ***MK2_ER*** | ***MK2_N*** |
| 1 | E1 | PER1 | S2 | DT1 | | 0.035 | 0.914 | 0.041 | 0.10 | 0.781 | 0.119 |
| 2 | | PER3 | | DT1 | | 0.177 | 0.738 | 0.075 | 0.012 | 0.80 | 0.188 |
| 3 | E2 | PER1 | S2 | | | 0.021 | 0.91 | 0.07 | 0.016 | 0.365 | 0.619 |
| 4 | E2 | | S1 | | ➡ | 0.00 | 0.349 | 0.651 | 0.15 | 0.831 | 0.019 |
| 5 | E2 | PER2 | | DT2 | | 0.07 | 0.03 | 0.90 | 0.177 | 0.301 | 0.522 |
| 6 | E3 | PER2 | S2 | DT1 | | 0.075 | 0.087 | 0.839 | 0.10 | 0.781 | 0.119 |
| 7 | | PER3 | S1 | DT2 | | 0.177 | 0.738 | 0.075 | 0.15 | 0.781 | 0.069 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

## Appendix C. Case Study

Table 27: Interrelation Tests for Asthma Data

| **Combination of two attributes (at $g = 2$)** | **Proportion of interrelated evidences ($\psi^p$)** | **Average of highest joint probability $\mu(p\|\psi)$** | **sparse index $\mathcal{S}$** |
|---|---|---|---|
| $\{DT, E\}$ | 1.0 | 0.62 | 0.896 |
| $\{DT, S\}$ | 0.916 | 0.58 | 0.915 |
| $\{DT, PE\}$ | 1.0 | 0.69 | 0.928 |
| $\{E, S\}$ | 0.916 | 0.63 | 0.925 |
| $\{E, PE\}$ | 1.0 | 0.79 | 0.909 |
| $\{S, PE\}$ | 0.916 | 0.77 | 0.928 |

Table 28: Interrelation Tests for Mortgage Loan

| **Combination of two attributes** | **Proportion of interrelated evidences ($\psi^p$)** | **Average of highest joint probability $\mu(p\|\psi)$** | **sparse index $\mathcal{S}$** |
|---|---|---|---|
| $\{A1, A2, A3, A4\}$ | 0.81 | 0.72 | 0.81 |
| $\{A5, A6\}$ | 0.89 | 0.78 | 1.0 |
| $\{A7, A8, A9, A10\}$ | 0.78 | 0.93 | 0.98 |
| $\{A11, A12\}$ | 0.88 | 0.78 | 0.77 |
| $\{A13, A14\}$ | 0.94 | 0.93 | 0.93 |
| $\{A15, A16, A17\}$ | 0.78 | 0.91 | 0.40 |

Table 29: Hyper-parameters of ANN for Early Asthma Symptoms Data

| | **Early Asthma Symptoms Data** | | |
|---|---|---|---|
| **Parameters** | **Scenario I** | **Scenario II** | **Scenario III** |

| | Missing proportion = 10.99% | | Missing proportion = 30.99% | | Missing proportion = 45.99% | |
|---|---|---|---|---|---|---|
| *Number of hidden layers (L)* | ER | 3 | ER | 4 | ER | 4 |
| | MAKER | 3 | MAKER | 4 | MAKER | 4 |
| | missForest | 3 | missForest | 4 | missForest | 4 |
| | MICE | 3 | MICE | 4 | MICE | 4 |
| | EM | 3 | EM | 4 | EM | 4 |
| | KNN | 3 | KNN | 4 | KNN | 4 |
| *Number of units per layer* | ER | $\{L_1 to\, L_2 = 12, L_3 = 2\}$ | ER | $\{L_1 to\, L_3 = 18, L_4 = 2\}$ | ER | $\{L_1 to\, L_3 = 18, L_4 = 2\}$ |
| | MAKER | $\{L_1 to\, L_2 = 9, L_3 = 2\}$ | MAKER | $\{L_1 to\, L_3 = 9, L_4 = 2\}$ | MAKER | $\{L_1 to\, L_3 = 9, L_4 = 2\}$ |
| | missForest | $\{L_1 to\, L_2 = 14, L_3 = 2\}$ | missForest | $\{L_1 to\, L_3 = 14, L_4 = 2\}$ | missForest | $\{L_1 to\, L_3 = 14, L_4 = 2\}$ |
| | MICE | $\{L_1 to\, L_2 = 14, L_3 = 2\}$ | MICE | $\{L_1 to\, L_3 = 14, L_4 = 2\}$ | MICE | $\{L_1 to\, L_3 = 14, L_4 = 2\}$ |
| | EM | $\{L_1 to\, L_2 = 14, L_3 = 2\}$ | EM | $\{L_1 to\, L_3 = 14, L_4 = 2\}$ | EM | $\{L_1 to\, L_3 = 14, L_4 = 2\}$ |
| | KNN | $\{L_1 to\, L_2 = 14, L_3 = 2\}$ | KNN | $\{L_1 to\, L_3 = 14, L_4 = 2\}$ | KNN | $\{L_1 to\, L_3 = 14, L_4 = 2\}$ |
| *Activation function* | ReLu: $L_1 to\, L_2$<br>SoftMax: $L_3$ | | ReLu: $L_1 to\, L_3$<br>SoftMax: $L_4$ | | ReLu: $L_1 to\, L_3$<br>SoftMax: $L_4$ | |
| *Dropout rate* | 20% at $L_3$ | | 20% at $L_4$ | | 10% at $L_4$ | |
| *Batch size* | 100 | | 100 | | 100 | |
| *Epoch* | 100 | | 100 | | 100 | |
| *Regularization strength* | $L^2$regularization strength = 0.01 in each layer | | $L^2$regularization strength = 0.01 in each layer | | $L^2$regularization strength = 0.01 in each layer | |
| *Learning rate* | 0.01 | | 0.001 | | .001 | |

*Layer ($L_1$) is input layer and number of units in first layer represents number of numerically transformed features. The data after imputation from missForest, MICE, EM, and KNN are transformed by to one-hot encode.

Table 30: Hyper-parameters of ANN for Mortgage Loan Data

| | **Mortgage Loan Data** | | | | | |
|---|---|---|---|---|---|---|
| **Parameters** | **Scenario I**<br>Missing proportion = 6.55% | | **Scenario II**<br>Missing proportion = 26.55% | | **Scenario III**<br>Missing proportion = 41.55% | |
| *Number of hidden layers (L)* | ER | 4 | ER | 5 | ER | 5 |
| | MAKER | 4 | MAKER | 5 | MAKER | 5 |
| | missForest | 4 | missForest | 5 | missForest | 6 |
| | MICE | 4 | MICE | 5 | MICE | 6 |
| | EM | 4 | EM | 5 | EM | 6 |
| | KNN | 4 | KNN | 5 | KNN | 6 |
| *Number of units per layer* | ER | $\{L_1 = 34, L_2\, to\, L_3 = 76, L_4 = 2\}$ | ER | $\{L_1 = 51, L_2\, to\, L_4 = 76, L_5 = 2\}$ | ER | $\{L_1 = 51, L_2\, to\, L_4 = 76, L_5 = 2\}$ |
| | MAKER | $\{L_1 = 12, L_2\, to\, L_3 = 25, L_4 = 2\}$ | MAKER | $\{L_1 = 18, L_2\, to\, L_3 = 30, L_4 = 25, L_5 = 2\}$ | MAKER | $\{L_1 = 18, L_2\, to\, L_4 = 35, L_5 = 2\}$ |
| | missForest | $\{L_1 = 60,$ | missForest | $\{L_1 = 60,$ | missForest | $\{L_1 = 60,$ |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | $L_2\ to\ L_3 = 80, L_4 = 2\}$ | | $L_2\ to\ L_4 = 80, L_5 = 2\}$ | | $L_2\ to\ L_4 = 80, L_5 = 70, L_6 = 2\}$ |
| | MICE | $\{L_1 = 60, L_2\ to\ L_3 = 80, L_4 = 2\}$ | MICE | $\{L_1 = 60, L_2\ to\ L_4 = 85, L_5 = 2\}$ | MICE | $\{L_1 = 60, L_2\ to\ L_4 = 80, L_5 = 75, L_6 = 2\}$ |
| | EM | $\{L_1 = 60, L_2\ to\ L_3 = 75, L_4 = 2\}$ | EM | $\{L_1 = 60, L_2\ to\ L_4 = 80, L_5 = 2\}$ | EM | $\{L_1 = 60, L_2\ to\ L_4 = 90, L_5 = 70, L_6 = 2\}$ |
| | KNN | $\{L_1 = 60, L_2\ to\ L_3 = 75, L_4 = 2\}$ | KNN | $\{L_1 = 60\ L_2\ to\ L_4 = 80, L_5 = 2\}$ | KNN | $\{L_1 = 60, L_2\ to\ L_4 = 85, L_5 = 70, L_6 = 2\}$ |
| *Activation function* | -*ReLu*: $L_1 to\ L_3$<br>-*SoftMax* in output layer | | -*ReLu*: $L_1 to\ L_4$<br>-*SoftMax* in output layer | | - *ReLu*: $L_1 to\ L_4$ for ER and MAKER. $L_1 to\ L_5$ for other data imputation methods<br>- *SoftMax* in output layer | |
| *Dropout rate* | 10% at $L_4$ | | 10% at $L_5$ | | 25% at $L_5$ | |
| *Batch size* | 100 | | 100 | | 100 | |
| *Epoch* | 100 | | 100 | | 100 | |
| *Regularization strength* | $L^2$regularization strength = 0.01 in each layer | | $L^2$regularization strength = 0.01 in each layer | | $L^2$regularization strength = 0.01 in each layer | |
| *Learning rate* | 0.001 | | 0.001 | | .002 | |

*Layer ($L_1$) is input layer and number of units in first layer represents number of numerically transformed features. The data after imputation from missForest, MICE, EM, and KNN are transformed by to one-hot encode.

Table 31: Hyper-parameters of Decision Tree for Early Asthma Symptoms Data

| | **Early Asthma Symptoms Data** | | | | | |
|---|---|---|---|---|---|---|
| **Parameters** | **Scenario I**<br>Missing proportion = 10.99% | | **Scenario II**<br>Missing proportion = 30.99% | | **Scenario III**<br>Missing proportion = 45.99% | |
| *Maximum depth of the tree* | ER | 5 | ER | 7 | ER | 7 |
| | MAKER | 5 | MAKER | 7 | MAKER | 9 |
| | missForest | 7 | missForest | 8 | missForest | 9 |
| | MICE | 8 | MICE | 8 | MICE | 8 |
| | EM | 8 | EM | 9 | EM | 11 |
| | KNN | 8 | KNN | 9 | KNN | 9 |
| *measure the quality of a split* | ER | gini | ER | gini | ER | gini |
| | MAKER | gini | MAKER | entropy | MAKER | entropy |
| | missForest | gini | missForest | gini | missForest | gini |
| | MICE | entropy | MICE | entropy | MICE | entropy |
| | EM | gini | EM | gini | EM | entropy |
| | KNN | entropy | KNN | gini | KNN | gini |
| *minimum number of samples to split node* | 2 | | 2 | | 2 | |

*default values were set for other parameters

Table 32: Hyper-parameters of Decision Tree for Mortgage Loan Data

| | **Mortgage Loan Data** | | | | | |
|---|---|---|---|---|---|---|
| **Parameters** | **Scenario I**<br>Missing proportion = 6.55% | | **Scenario II**<br>Missing proportion = 26.55% | | **Scenario III**<br>Missing proportion = 41.55% | |
| *Maximum depth of the tree* | ER | 7 | ER | 9 | ER | 11 |
| | MAKER | 7 | MAKER | 8 | MAKER | 11 |
| | missForest | 8 | missForest | 11 | missForest | 13 |
| | MICE | 9 | MICE | 12 | MICE | 13 |
| | EM | 9 | EM | 13 | EM | 14 |
| | KNN | 9 | KNN | 12 | KNN | 13 |
| *measure the quality of a split* | ER | gini | ER | gini | ER | entropy |
| | MAKER | gini | MAKER | gini | MAKER | gini |
| | missForest | gini | missForest | gini | missForest | gini |
| | MICE | gini | MICE | entropy | MICE | entropy |
| | EM | gini | EM | gini | EM | entropy |
| | KNN | entropy | KNN | gini | KNN | gini |
| *minimum number of samples to split node* | 2 | | 2 | | 2 | |

*default values were set for other parameters

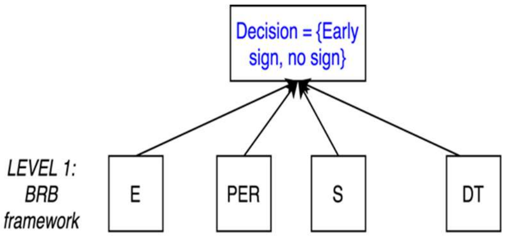


Figure 15. BRB structure for asthma symptoms, individual input attributes are transformed by I-MAKER

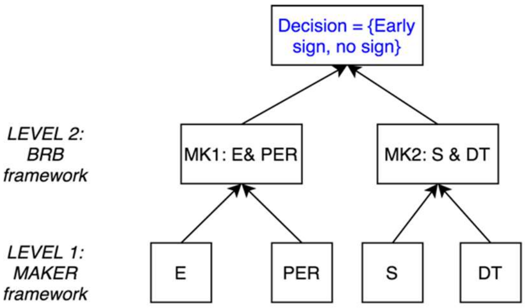


Figure 16. BRB structure for asthma symptoms, input attributes are combined and transformed by C-MAKER

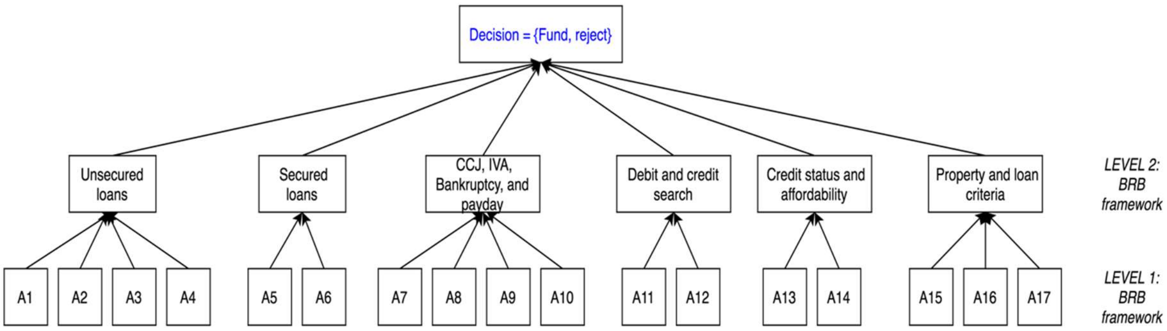


Figure 17. BRB structure for mortgage loan, individual input attributes are transformed by I-MAKER

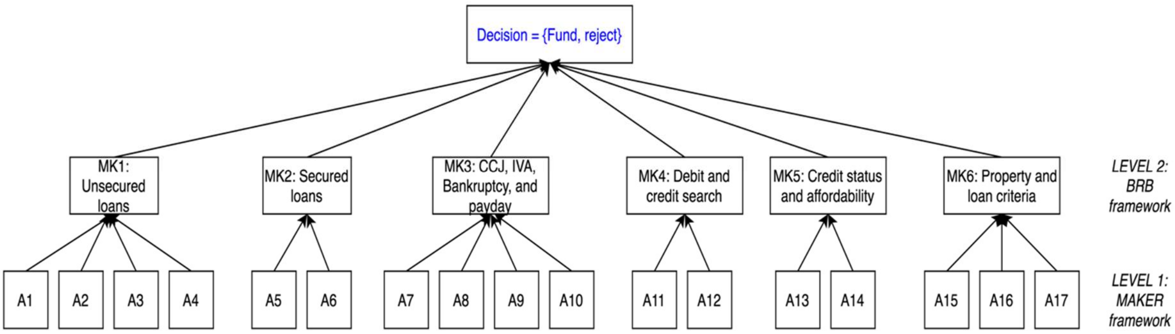

Decision = {Fund, reject}
MK1: Unsecured loans
MK2: Secured loans
MK3: CCJ, IVA, Bankruptcy, and payday
MK4: Debit and credit search
MK5: Credit status and affordability
MK6: Property and loan criteria
LEVEL 2: BRB framework
A1
A2
A3
A4
A5
A6
A7
A8
A9
A10
A11
A12
A13
A14
A15
A16
A17
LEVEL 1: MAKER framework

Figure 18. BRB structure for mortgage loan, input attributes are combined and transformed by C-MAKER