IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 15, 2020, accepted June 29, 2020, date of publication July 10, 2020, date of current version July 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3008475

Using Partial Combination Models to Improve Prediction Quality and Transparency in Mixed Datasets

YI-HSIN WU[1,2], (Associate Member, IEEE), YU-HSIN CHANG[3], YIN-JING TIEN[2], CHENG-JUEI YU[2], SHENG-DE WANG[1], (Member, IEEE), AND CHENG-HUNG WU[3], (Member, IEEE)

[1]Department of Electrical Engineering, National Taiwan University, Taipei 106319, Taiwan
[2]Institute for Information Industry, Taipei 106214, Taiwan
[3]Institute of Industrial Engineering, National Taiwan University, Taipei 106319, Taiwan

Corresponding author: Cheng-Hung Wu (wuchn@ntu.edu.tw)

This work was supported in part by the Advanced Artificial Intelligence Technologies and Industry Applications (3/4) under Grant 109-EC-17-A-21-1516 and in part by III Innovative and Prospective Technologies Project (1/1) under Grant 109-EC-17-A-24-0461, which are subsidized by the Ministry of Economic Affairs in Taiwan. The work of the Cheng-Hung Wu was supported in part by the Ministry of Science and Technology, Taiwan under Grant MOST107-2628-E-002-006-MY3.

ABSTRACT Mixed Datasets with complex interactions between categorical and numerical attributes are common in engineering and business applications. For example, production rates in manufacturing systems are jointly influenced by several categorical and numerical attributes, such as machine and product types and their numerical attributes. This study aims to improve the prediction performance and transparency of mixed datasets with complex interactions using machine learning (ML) methods. The proposed method requires lesser data and computational effort than existing hierarchical or clustering regression methods. Multiple prediction models can be generated by partitioning a dataset into subsets with different categorical attribution combinations. One- and two-stage model selection methods are proposed to use the training and validation datasets in selecting better models among all the prediction models. Numerical results demonstrate the potential of the model selection approach in a mixed dataset from a semiconductor manufacturer. In comparison with regression models, more than 30% reduction in root mean square error is observed using the proposed model selection approach. The cross-validation test results also demonstrated a 10% improvement in accuracy against the properly tuned XGBoost models. Moreover, the proposed model selection approach is compatible with other regression or ML prediction methods and can be used to improve the model's transparency of any existing methods on mixed datasets.

INDEX TERMS Hierarchical method, hierarchical clustering, prediction methods, regression analysis, manufacturing, expert systems.

I. INTRODUCTION

Mixed Datasets with complex interactions between categorical and numerical attributes are common in engineering and business applications. For example, production rates in manufacturing systems are jointly influenced by several categorical and numerical attributes, such as machine and product types and their numerical attributes. This study aims to improve the prediction performance for mixed datasets with complex interactions using machine learning (ML) methods.

Prediction quality is crucial for operation efficiency of manufacturing and service industries. For example, predicting the throughput rate of a specific machine–product combination is a critical task for scheduling, capacity planning, and other operations management activities in manufacturing industry. The proposed methodology helps practitioners improve the overall prediction accuracy and provide transparency of the prediction models. As indicated in [1], tool cost consists of 70% of the total costs of semiconductor manufacturing. Underestimation or overestimation of the throughput rate will lead to surplus or shortage of capacity. In capital intensive industries, such as semiconductor manufacturing, companies annually invest billions in capacity expansion, and the proposed method can significantly enhance decision quality.

This study is driven by the need to predict the production rate of new products among multiple machine types in semiconductor testing and assembly facilities. Fig. 1 shows the

The associate editor coordinating the review of this manuscript and approving it for publication was Xiao-Sheng Si.

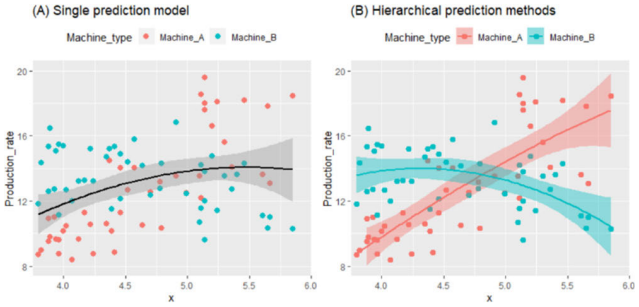


FIGURE 1. Production rates of two machine types: (A) without and (B) with the consideration of the interaction between machine types and x-attributes.

production rates of different products on two machines, wherein the x- and y-axes represent the numerical attribute of the products and the production rates, respectively. The data in Fig. 1 shows the clear interactions between the machine type and the x-value, which suggests that the red data from Machine A and the blue data from Machine B behave differently. In this example, building a single prediction model for both machines without considering the interactions between the machine types and x-attributes becomes unsatisfactory (Fig. 1 [A]). Thus, hierarchical prediction methods are necessary in such environments (Fig. 1[B]).

Given that capacity expansions are often applied over time gradually, different machine types can be procured at various capacity expansion stages. Thus, machines types can be a common categorical attribute in manufacturing systems.

Moreover, many products can be produced simultaneously in low volumes when product life cycles detract and additional product customizations are adopted. In such low-volume and high-mix (LVHM) manufacturing systems, product types and the other categorical product attributes must also be considered.

While multiple machine types and LVHM products are simultaneously present in a manufacturing system, production rates or yield qualities may be jointly influenced by the complex interactions between the categorical and numerical attributes of the machines and products. Thus, modelling the complex interactions among attributes is crucial for improving decisions regarding scheduling or dispatching. However, in LVHM environments with short product life cycle and multiple machine types, new products and product–machine combinations that lack historical production rate or quality data are common and cause difficulties in production planning.

To overcome the aforementioned challenges in LVHM production systems, the main contributions of this study are summarized as follows. Fig. 2 shows the general framework for using partial combination models to improve prediction quality and transparency.

- This research is among the first studies to adopt the model selection approach for prediction.
- The proposed method improves prediction quality. In comparison with the popular XGBoost models or other commercial packages, the overall root mean square error (RMSE) can be reduced by more than 10%.
- The proposed method improves the model's transparency.
- The proposed method requires lesser data and computational effort than hierarchical regression.
- The proposed method is robust in noisy environments, especially when outliers exist.

The remainder of the work is arranged as follows. Section II reviews related works. Sections III and IV formally defines the proposed prediction models. Section V presents the model selection methods. Section VI uses the empirical dataset from a semiconductor manufacturer to validate the overall performance of the proposed method. Section VII summarizes the findings and provides insights into the general mixed dataset prediction problems.

II. LITERATURE REVIEW

In literature, machine learning methods are commonly used to predict a numerical response variable in mixed datasets. In the semiconductor manufacturing domain, [2] compared four machine learning methods that can handle mixed datasets to estimate factory cycle time, and found that decision tree regression method has the best prediction performance. Reference [3] developed a tree-based piecewise linear regression model to estimate the flow-time of a manufacturing system. Reference [4] used different machine learning approaches to improve the lead time prediction for a mixed dataset from a manufacturing execution system. Tree-based ensemble methods have the lowest root mean square error (RMSE) and mean absolute error (MAE). References [5] and [6] used XGBoost tree-based classifier for the mixed datasets from the Kaggle competition, ''Bosch Production Line Performance''. Reference [7] used XGBoost

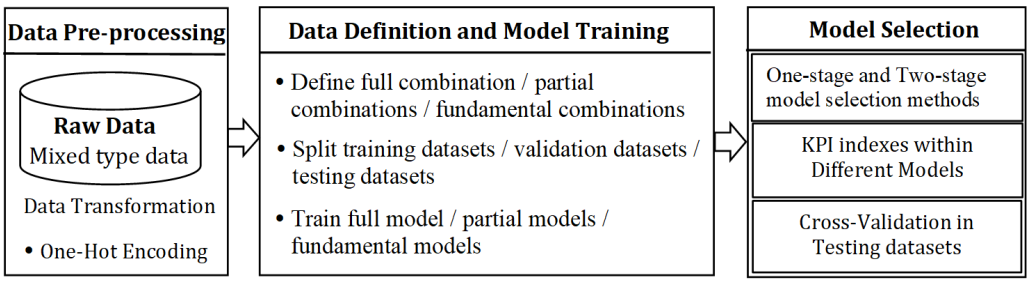


FIGURE 2. The general framework for using partial combination models.

method to improve prediction accuracy in intrusion detection for mixed datasets. Among the relevant research, XGBoost and decision tree regression are commonly recognized as the best performing prediction methods for mixed datasets. Thus, these two methods are used as benchmarks for our numerical study to show the superiority of the proposed algorithm.

Further literature review includes four machine learning approaches and concepts, namely, hierarchical regression/clustering methods, feature selection, ensemble methods, and model transparency. These approaches are widely used and discussed for prediction problems in mixed datasets.

A. HIERARCHICAL REGRESSION/CLUSTERING METHODS

Performing data clustering followed by the application of regression methods on each cluster are often suggested when conducting regression analysis on a dataset with complex interaction among variables. In polynomial regression, the dependent variable y is modeled as an n-th degree polynomial of independent variable x [13]. Segmented regression is a classical statistical analysis method that uses this philosophy [8]. The method constructs piecewise linear (or nonlinear) functions for the different portions of a dataset. Despite the restriction in cluster shape, other clustering methods, including agglomerative, density based spatial clustering of applications with noise (DBSCAN), Gaussian mixtures, and k-means clustering, are designed for purely numerical datasets because the use of binary categorical variables (e.g., 1 and 0) can distort distance computation even if the data is normalized. Gaussian mixtures, which assume that the dataset forms a mixture of a finite number of small distributed Gaussian datasets, may not be scalable to large datasets. Agglomerative clustering, which belongs in the family of hierarchical clustering methods, builds clusters by pairing up clusters according to similarity. Agglomerative clustering uses a bottom–up approach that begins from individual data points to a certain threshold [14]. DBSCAN is a density based algorithm used for discovering clusters. This algorithm can capture uneven-shaped clusters, but its computation performance deteriorates with the dimensionality of the feature space [15].

Clustering methods require large computational effort for Mixed Datasets that contain numerous features. In addition, categorical variables severely affect the determination of the similarity among data points even after pre-processing (e.g., scaling and normalization). Moreover, when using clustering for regression, considering the regression performance (e.g., R^2, mean square error) instead of only the common clustering performance evaluation metrics (e.g., similarity measures) remains a major challenge.

Clustering methods are useful on datasets with mixed categorical and numerical variables. Reference [9] adopted the clustering concept into the clustered linear regression, which improved the performance of the classical linear regression by determining the partitions that enhanced the accuracy of local linear regressions. Similarly, multilevel or hierarchical regression allows different regression parameters for each portion/cluster of the dataset by inputting the relationships among different variables (i.e., hierarchies) [10]. Hierarchical linear regression is another widely used method for mixed data sets. By adding or deleting variables, hierarchical linear regression finds independent variables that have significant influence on the response variable [16]. Hierarchical linear modeling (HLM) is a useful regression method for mixed dataset with hierarchical relationship among predictive variables [17]. However, complex mathematical calculations and statistical techniques are required in establishing HLM models. In such models, the number of model parameters exponentially increases with the number of levels or attributes; thus, HLM has an unsatisfactory performance in datasets with large number of attributes/levels.

Although the early clustering methods fail to perform well in modern applications with large number of features, the clustering concept is widely adopted in contemporary ML and hierarchical regression models, especially in the unsupervised learning paradigm. In predicting a numerical response variable, a method that guides clustering through regression functions can still be beneficial. Spath's regression exchange algorithm improves the regression performance of clusters by exchanging data points between clusters [18]. [11] provided the term regression clustering for the family of cluster-wise regression methods by adopting the regression exchange algorithm. In regression clustering, the dataset is partitioned into clusters using a center-based clustering algorithm (e.g., k-means clustering) through the help of regression optimization algorithms. In these methods, clustering is applied iteratively, and the objective function and stopping criterion are based on regression performance. Reference [12] developed algorithms for generalized cluster-wise linear regression problems. TABLE 1 summarizes different hierarchical and clustering methods and their corresponding limitations.

B. FEATURE SELECTION

A common limitation of hierarchical or clustering methods is the exponential growth of model parameters along with categorical attributes. To overcome this limitation, feature selection is used in traditional statistical analyses for dimension reduction. Feature selection selects available features by eliminating irrelevant ones. Some works have proposed methods for dimension reduction when dealing with large data [19], [20]. A new feature selection method for crisp and low-quality data was proposed in [21].

Although feature selection methods are commonly used, they are not beneficial when all features or attributes are relevant. For instance, in Fig. 1, categorical and numerical attributes are relevant, and feature selection methods will not reduce the dimension of the regression problem. Moreover, in some large prediction problems, the dimension of the regression problem remains excessively high for efficient computation. Many studies used ML methods to improve the prediction performance for overcoming the aforementioned disadvantages of traditional statistical methods.

TABLE 1. Hierarchical and clustering-based regression methods.

Method	Description	Limitation
Segmented regression [8]	Fits piecewise linear (or non-linear) functions	Limited proximity and low dimensions
Clustered linear regression [9]	Performs linear regression on clusters	Clustering and regression are performed and evaluated separately. Limited proximity and low dimensions
Hierarchical methods [10]	Nested/multilevel regression model based on the relationships of variables	High computational complexity unless supervised by analyst
Regression clustering [11]	Performs clustering and regression iteratively	Only applicable to purely numerical datasets
Cluster-wise linear regression [12]	Performs clustering based on the overall sum of the squared errors	High computational complexity. Fails to consider complex interaction among attributes

C. ENSEMBLE METHODS

For manufacturing applications, many researchers have constructed prediction models using ensemble methods [22]–[24]. Decision tree regression is an ensemble method using tree structures to handle numerical and categorical data. Decision tree regression needs complex calculation and more time to train models than other ML methods [13]. References [25] and [26] apply ensemble methods to analyze semiconductor process data. Some works utilized ensemble approaches for intrusion detection [27]. Compared with traditional statistical methods, ensemble methods can transform the complex and non-linear characteristics of data effectively and train models to improve the efficiency of classification and prediction. Gradient boosting is a practical and popular ensemble method. As an open-source implementation of gradient boosting methods, XGBoost became prominent in ML competitions and data mining challenges (e.g., ML competitions held by Kaggle). For example, in the KDDCup 2015 and ICDN challenge 2015, XGBoost demonstrated a remarkable performance over a wide range of data classification problems. Reference [28] detailed XGBoost, which was a scalable tree boosting system that was favored by many ML competition winning teams because of its high performance and computational speed. This system performed model adaptation with high flexibility and produced state-of-the-art model results. Many studies applied extreme gradient boosting methods to different ML regions. The system yielded a comparatively better performance than the original one by constructing an XGBoost-based prediction model for short-term load prediction [29]. Reference [30] developed an XGBoost framework for biomedical fields to predict essential proteins. In addition, [31] used the XGBoost algorithm to classify patients with focal epilepsy. However, ensemble methods, such as the XGBoost framework, still suffer from model transparency issues. To overcome such issues, this study improves prediction quality based on the satisfactory performance of modern ML packages. The proposed method initially constructs a large number of models under different combinations of categorical attributes and then develops a model selection method that will select among the prediction models to improve the overall prediction quality and the model's transparency.

D. MODEL TRANSPARENCY

The increasingly complex prediction models and machine learning methods have led to the concerns of model transparency. The correlation and logic between models are established by proposing interpretable classifiers on the bases of statistical probability [32]. Reference [33] develops interpretable decision sets and uses independent if-then rules to build interpretable models. A novel tree model splitting criterion is proposed to enhance model interpretability [34]. Other model transparency related research is summarized in the survey of [35]. However, despite the increasing attention on model transparency, the hidden correlation between attributes cannot be sufficiently explained through existing ML methods, and most machine learning methods still focus on model accuracy improvement. In this research, the use of the partial combination data set improves model transparency through providing insights into the hidden correlation between categorical attributes.

Although numerous research studies mixed datasets for prediction, entire datasets are used to build a single prediction model that may not consider the complex interactions between categorical variables. In addition, model transparency is commonly neglected. Thus, our proposed method splits datasets for building models and selects the best prediction models to improve the prediction accuracy and model transparency.

III. PROBLEM DESCRIPTION AND DATA PREPROCESSING

We assume a mixed data set (Γ) consisting of k categorical explanatory variables ($X_i, i = 1 \ldots .k$), m numerical explanatory variables ($X_l, l = k{+}1 \ldots k + m$), and one numerical response variable (Y). Each categorical variable ($X_i, i = 1 \ldots k$) has ($N_i, i = 1 \ldots k$) distinct values. Let $\Omega_i = \{1, 2, 3 \ldots \ldots N_i\}, i = 1 \ldots k$ denote the sample space for the values of the i^{th} categorical attribute; $x_i \in \Omega_i$ is a specific value for the i^{th} categorical explanatory variables ($X_i, i = 1 \ldots k$). The complex interaction effects occur when the categorical and/or numerical variables interact with each other.

TABLE 2. Mixed dataset (Γ) considered in this research.

MIXED DATASET(Γ)								
	Explanatory variable							Response variable
Type	Categorical				Numerical			Numerical
Variables	X_1	X_2	...	X_k	X_{k+1}	...	X_{k+m}	Y
Value of the variable	$x_1 \in \Omega_1$	$x_2 \in \Omega_2$	...	$x_k \in \Omega_k$	$x_{k+1}, x_{k+2}, \ldots\ldots, x_{k+m} \in R$			$y \in R$
Observation 1	x_1^1	x_2^1	...	x_k^1	x_{k+1}^1	...	x_{k+m}^1	y^1
Observation 2	x_1^2	x_2^2	...	x_k^2	x_{k+1}^2	...	x_{k+m}^2	y^2
...	...	...	...	...	...	...	...	...
Observation P	x_1^P	x_2^P	...	x_k^P	x_{k+1}^P	...	x_{k+m}^P	y^P

In summary, in TABLE 2, we assume that a dataset (Γ) contains p independent observations, each with k categorical features, m numerical features, and a numerical outcome. Let $\left(x_1^p, x_2^p, x_3^p, \ldots, x_k^p, x_{k+1}^p, \ldots, x_{k+m}^p, y^p\right)$, $p = 1 \ldots P$ be the p-th observation of Γ, then Γ can be defined as a $[P] \times [k + m + 1]$ matrix (e.g., $\left(x^2, y^2\right) = \left(x_1^2, x_2^2, x_3^2, \ldots, x_k^2, x_{k+1}^2, \ldots, x_{k+m}^2, y^2\right)$ which is the second observation in the dataset).

In the Mixed Dataset Γ, subsets are defined according to the categorical attribute values of each observation, beginning from the fundamental dataset, which contains only the observations with identical categorical attributes. Then, partial and full combination datasets are defined through the union of fundamental datasets.

A. FUNDAMENTAL COMBINATION $(x_1, x_2, \ldots\ldots x_k)$ OF CATEGORICAL ATTRIBUTES AND DATASETS

1) FUNDAMENTAL COMBINATION $(x_1, x_2, \ldots\ldots x_k)$

Let $(x_1, x_2, \ldots\ldots x_k)$ be a specific combination of the categorical attributes, where $(x_1, x_2, \ldots\ldots x_k)$ is the fundamental combination of the categorical attributes or simply the fundamental combination. Given that the categorical attribute X_i has N_i possible values, $\prod_{i=1}^{k} N_i$ fundamental combinations can be obtained at most.

2) FUNDAMENTAL COMBINATION DATASET $\Gamma_{x_1, x_2, \ldots x_k}$

Let $\Gamma_{x_1, x_2, \ldots x_k} \subset \Gamma$ be a set that contains all observations that satisfy $(X_1, X_2, \ldots\ldots X_k) = (x_1, x_2, \ldots\ldots x_k)$ (i.e., the categorical attributes are identical to the fundamental combination). Different fundamental combination datasets are mutually exclusive. The union of all fundamental combination datasets is the entire dataset Γ.

3) THREE MUTUALLY EXCLUSIVE SUBSETS $\Gamma^{Tr}_{x_1, x_2, \ldots x_k}$, $\Gamma^{V}_{x_1, x_2, \ldots x_k}$, AND $\Gamma^{Ts}_{x_1, x_2, \ldots x_k}$ OF THE FUNDAMENTAL COMBINATION DATASETS $\Gamma_{x_1, x_2, \ldots x_k}$, WHERE $\Gamma_{x_1, x_2, \ldots x_k} = \Gamma^{Tr}_{x_1, x_2, \ldots x_k} \cup \Gamma^{V}_{x_1, x_2, \ldots x_k} \cup \Gamma^{Ts}_{x_1, x_2, \ldots x_k}$

- $\Gamma^{Tr}_{x_1, x_2, \ldots x_k}$: Fundamental combination training dataset, which contains 70% randomly selected data from $\Gamma_{x_1, x_2, \ldots x_k}$; $\Gamma^{Tr}_{x_1, x_2, \ldots x_k}$ is used for the model training.
- $\Gamma^{V}_{x_1, x_2, \ldots x_k}$: Fundamental combination validation dataset, which contains 15% randomly selected data from $\Gamma_{x_1, x_2, \ldots x_k}$; $\Gamma^{V}_{x_1, x_2, \ldots x_k}$ is used for model validation and selection, which will be discussed in the next section.
- $\Gamma^{Ts}_{x_1, x_2, \ldots x_k}$: Fundamental combination testing dataset, which contains 15% randomly selected data from $\Gamma_{x_1, x_2, \ldots x_k}$. The testing dataset is used to measure the performance of the proposed method.

The functions of these subsets will be explained in the model's training and selection sections.

B. PARTIAL COMBINATION OF CATEGORICAL ATTRIBUTES AND PARTIAL COMBINATION DATASETS

Let I be a subset of $\{1, 2, \ldots k\}$, $I \neq \emptyset$ and $I \neq \{1, 2, \ldots k\}$.

1) PARTIAL COMBINATION $(x_j, j \notin I)$

Let $(x_j, j \notin I)$ be a partial combination of the categorical attributes, in which X_j can be an arbitrary value in Ω_j when $j \in I$, otherwise, X_j should be a specific value x_j in Ω_j. $(x_j, j \notin I)$ assigns specific values only to a part of the categorical attributes, hence, $(x_j, j \notin I)$ is a partial combination of the categorical attributes or simply partial combination.

2) PARTIAL COMBINATION DATASET $\Gamma_{(x_j, j \notin I)}$

Let $\Gamma_{(x_j, j \notin I)} = \bigcup_{[x_i \in \Omega_i, i \in I]} \Gamma_{x_1, x_2, \ldots x_k}$ be the union of all fundamental combination datasets that satisfies the partial combination (i.e., $\Gamma_{(x_j, j \notin I)}$ is a set that contains all data that satisfy the partial combination of the categorical attributes and $\left(X_j = x_j, j \notin I\right)$). Note that different partial combination datasets might not be mutually exclusive and might be identical.

3) THREE MUTUALLY EXCLUSIVE SUBSETS $\Gamma^{Tr}_{(x_j, j \notin I)}$, $\Gamma^{V}_{(x_j, j \notin I)}$, AND $\Gamma^{Ts}_{(x_j, j \notin I)}$ OF THE PARTIAL COMBINATION DATASETS $\Gamma_{(x_j, j \notin I)}$, WHERE $\Gamma_{(x_j, j \notin I)} = \Gamma^{Tr}_{(x_j, j \notin I)} \cup \Gamma^{V}_{(x_j, j \notin I)} \cup \Gamma^{Ts}_{(x_j, j \notin I)}$

- $\Gamma^{Tr}_{(x_j, j \notin I)} = \bigcup_{[x_i \in \Omega_i, i \in I]} \Gamma^{Tr}_{x_1, x_2, \ldots x_k}$: Partial combination training dataset

- $\Gamma^{V}_{(x_j,j\notin I)} = \bigcup_{[x_i\in\Omega_i,i\in I]} \Gamma^{V}_{x_1,x_2,\ldots x_k}$: Partial combination validation dataset
- $\Gamma^{Ts}_{(x_j,j\notin I)} = \bigcup_{[x_i\in\Omega_i,i\in I]} \Gamma^{Ts}_{x_1,x_2,\ldots x_k}$: Partial combination testing dataset

$\Gamma^{Tr}_{(x_j,j\notin I)}$, $\Gamma^{V}_{(x_j,j\notin I)}$, and $\Gamma^{Ts}_{(x_j,j\notin I)}$ still contain 70%, 15%, and 15% of the data in the entire partial combination dataset $\Gamma_{(x_j,j\notin I)}$, respectively; and all training, validation, and testing data in the fundamental datasets assume the respective similar functions in the partial combination datasets. None of the testing data will be used for the training of any models in the next section.

C. FULL COMBINATION OF CATEGORICAL ATTRIBUTES AND THEIR CORRESPONDING DATASETS

Following the definition of the partial combination, when $I = \{1, 2, \ldots k\}$, all categorical attributes (X_j) can be arbitrary values in Ω_j.

1) FULL COMBINATION $(\Omega_1, \Omega_2, \ldots \Omega_k)$

Let $(\Omega_1, \Omega_2, \ldots \Omega_k)$ be the full combination of the categorical attributes, where such attributes can all be arbitrary.

2) FULL COMBINATION DATASET $\Gamma_{\Omega_1,\Omega_2,\ldots\Omega_k} = \bigcup_{[x_i\in\Omega_i,i\in I]} \Gamma_{x_1,x_2,\ldots x_k}$

Let $\Gamma_{\Omega_1,\Omega_2,\ldots\Omega_k} = \bigcup_{[x_i\in\Omega_i,i\in I]} \Gamma_{x_1,x_2,\ldots x_k}$ be the union of all fundamental combination datasets. Thus, $\Gamma_{\Omega_1,\Omega_2,\ldots\Omega_k}$ is also the entire dataset, and $\Gamma_{\Omega_1,\Omega_2,\ldots\Omega_k} = \Gamma$.

3) THREE MUTUALLY EXCLUSIVE SUBSETS $\Gamma^{Tr}_{\Omega_1,\Omega_2,\ldots\Omega_k}$, $\Gamma^{V}_{\Omega_1,\Omega_2,\ldots\Omega_k}$, AND $\Gamma^{Ts}_{\Omega_1,\Omega_2,\ldots\Omega_k}$ OF THE FULL COMBINATION DATASET $\Gamma_{\Omega_1,\Omega_2,\ldots\Omega_k}$

Let $\Gamma_{\Omega_1,\Omega_2,\ldots\Omega_k} = \Gamma^{Tr}_{\Omega_1,\Omega_2,\ldots\Omega_k} \cup \Gamma^{V}_{\Omega_1,\Omega_2,\ldots\Omega_k} \cup \Gamma^{Ts}_{\Omega_1,\Omega_2,\ldots\Omega_k}$, where $\Gamma_{\Omega_1,\Omega_2,\ldots\Omega_k}$ represents the entire dataset Γ because $\Gamma_{\Omega_1,\Omega_2,\ldots\Omega_k}$ includes the observations of arbitrary categorical variable values. For simplicity, $\Gamma^{Tr} = \Gamma^{Tr}_{\Omega_1,\Omega_2,\ldots\Omega_k}$, $\Gamma^{V} = \Gamma^{V}_{\Omega_1,\Omega_2,\ldots\Omega_k}$, and $\Gamma^{Ts} = \Gamma^{Ts}_{\Omega_1,\Omega_2,\ldots\Omega_k}$ are respectively defined as the full training, validation, and testing datasets.

- $\Gamma^{Tr} = \Gamma^{Tr}_{\Omega_1,\Omega_2,\ldots\Omega_k} = \bigcup_{[x_i\in\Omega_i,i\in I]} \Gamma^{Tr}_{x_1,x_2,\ldots x_k}$: Full combination training dataset
- $\Gamma^{V} = \Gamma^{V}_{\Omega_1,\Omega_2,\ldots\Omega_k} = \bigcup_{[x_i\in\Omega_i,i\in I]} \Gamma^{V}_{x_1,x_2,\ldots x_k}$: Full combination validation dataset
- $\Gamma^{Ts} = \Gamma^{Ts}_{\Omega_1,\Omega_2,\ldots\Omega_k} = \bigcup_{[x_i\in\Omega_i,i\in I]} \Gamma^{Ts}_{x_1,x_2,\ldots x_k}$: Full combination testing dataset

$\Gamma^{Tr}_{\Omega_1,\Omega_2,\ldots\Omega_k}$, $\Gamma^{V}_{\Omega_1,\Omega_2,\ldots\Omega_k}$, and $\Gamma^{Ts}_{\Omega_1,\Omega_2,\ldots\Omega_k}$ contain 70%, 15%, and 15% of the data in the entire dataset $\Gamma_{\Omega_1,\Omega_2,\ldots\Omega_k}$, respectively; and all training, validation, and testing data in the fundamental datasets maintain the same functions in $\Gamma_{\Omega_1,\Omega_2,\ldots\Omega_k}$. None of the testing data will be used for the training of any models in the next section.

IV. FUNDAMENTAL, PARTIAL, AND FULL COMBINATION PREDICTION MODELS USING THE CORRESPONDING DATASETS

A distinct prediction model will be trained for every distinct fundamental, partial, and full combination training dataset $\mathbf{\Gamma^{Tr}_{x_1,x_2,\ldots x_k}}$, $\mathbf{\Gamma^{Tr}_{(x_j,j\notin I)}}$, and $\mathbf{\Gamma^{Tr}_{\Omega_1,\Omega_2,\ldots\Omega_k}}$. Model training can be conducted using any statistical or ML methods. In the numerical study of this research, one-hot encoding is applied to the categorical attributes and XGBoost is used for the model training in the numerical analysis.

A. FUNDAMENTAL COMBINATION PREDICTION MODEL $(M_{x_1,x_2,\ldots x_k})$

Let $M_{x_1,x_2,\ldots x_k}$ be the prediction model trained by the fundamental training dataset $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$; $M_{x_1,x_2,\ldots x_k}$ is the fundamental combination prediction model or simply the fundamental model.

B. PARTIAL COMBINATION PREDICTION MODEL $(M_{(x_j,j\notin I)})$

Let $M_{(x_j,j\notin I)}$ be the prediction model trained by the partial combination training dataset $\Gamma^{Tr}_{(x_j,j\notin I)}$, when $I \neq \emptyset$ and $I \neq \{1, 2 \ldots k\}$; $M_{(x_j,j\notin I)}$ is the partial combination prediction model or the partial model.

C. FULL COMBINATION PREDICTION MODEL $(M_{\Omega_1,\Omega_2,\ldots,\Omega_k})$

Let $M_{\Omega_1,\Omega_2,\ldots,\Omega_k}$ be the prediction model trained by the full combination training dataset $\Gamma^{Tr}_{\Omega_1,\Omega_2,\ldots\Omega_k}$; $M_{\Omega_1,\Omega_2,\ldots,\Omega_k}$ is the full combination prediction model or the full model.

Note that none of the validation and testing data is used for the model training in all fundamental, partial, and full prediction models because $\Gamma^{Tr}_{\Omega_1,\Omega_2,\ldots\Omega_k}$ and $\Gamma^{Tr}_{(x_j,j\notin I)}$ are generated from the union of the fundamental training dataset $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$; all the fundamental datasets are mutually exclusive. This condition ensures the quality of the validation and testing processes, which will be introduced in the subsequent section.

V. MODEL TRAINING PROCESS

For the training process adopted in the proposed algorithm, Table 3 summarizes attributes used for training different models. Each fundamental dataset $(\Gamma_{x_1,x_2,\ldots x_k})$ has the same specific values for all categorical attributes $(X_1, X_2, \ldots, X_k) = (x_1, x_2, \ldots, x_k)$. Thus, given that the same values of categorical attributes do not offer any useful information for predicting the response, only the numerical attributes are used to train fundamental models $M_{x_1,x_2,\ldots x_k}$. Then, combining different fundamental datasets with similar prediction models allows more observations to be used for better estimation of model parameters of the common prediction model and improve its accuracy. For the partial combination model, we calculate the union of corresponding fundamental datasets to obtain each partial dataset $\Gamma_{(x_j,j\notin I)} = \bigcup_{[x_i\in\Omega_i,i\in I]} \Gamma_{x_1,x_2,\ldots x_k}$, $I \neq \emptyset$ and $I \neq \{1, 2, \ldots k\}$, in which all of the data in $\Gamma_{(x_j,j\notin I|I\neq\emptyset \text{ and } I\neq\{1,2\ldots k\})}$ have the same specific values for categorical attributes $i\notin I$.(i.e., $X_i = x_j$, $i \notin I$.) Again, given that the same values of categorical attributes do not offer any useful information for predicting the response, only categorical attributes X_i, $i \in I$ with different values are used to train the corresponding partial combination model $M_{(x_j,j\notin I|I\neq\emptyset \text{ and } I\neq\{1,2\ldots k\})}$. In addition, given that the full

TABLE 3. Attributes used for training different models.

Model	Full models $M_{\Omega_1,\Omega_2,\ldots,\Omega_k}$	Partial models $M_{(x_j, j\notin I \mid I\neq\emptyset \text{ and } I\neq\{1,2\ldots k\})}$	Fundamental models $M_{x_1,x_2,\ldots x_k}$
Categorical attributes	All	$X_i, i \in I$	None
Numerical attributes	All		

dataset ($\Gamma_{\Omega_1,\Omega_2,\ldots\Omega_k} = \bigcup_{[x_i\in\Omega_i, i\in I]} \Gamma_{x_1,x_2,\ldots x_k}$) represents the union of all fundamental datasets and includes all combinations of categorical variables, all categorical and numerical attributes are used to train the full combination model $M_{\Omega_1,\Omega_2,\ldots,\Omega_k}$.

Using the corresponding datasets with all numerical attributes and the above chosen categorical attributes, we train all fundamental models ($M_{x_1,x_2,\ldots x_k}$), partial models ($M_{(x_j,j\notin I\mid I\neq\emptyset \text{ and } I\neq\{1,2\ldots k\})}$) and full models ($M_{\Omega_1,\Omega_2,\ldots,\Omega_k}$) using XGBoost. Grid search in [36] is used for finding the best values of hyper-parameters (general, booster, and learning task parameters) in the training. In summary, all features are still used in the full combination model and in several partial combination models. Thus, none of the features are completely removed throughout the model selection approach unlike most feature selection methods.

More precisely, the model selection finds similarities between different fundamental datasets and then selects the best prediction model for each fundamental combination. Consider the semiconductor dataset [37] used in our numerical study, all the attributes are important to predict the response variable and none can be completely removed from the prediction models. However, several categorical variable combinations may have similar regression models and pooling those fundamental datasets enhances the estimation of the prediction model parameters. Without using the proposed model selection prediction approach, pooling together fundamental or partial combination datasets without similarities hampers the model parameter estimations, leading to negative impacts on the overall prediction accuracy.

Each fundamental dataset is included in exactly 2^k prediction models, including one fundamental prediction model, 2^k-2 partial combination prediction models, and one full combination model. Given that each categorical attribute can either be specific or arbitrary in the partial datasets, 2^k datasets are associated with each fundamental combination. Excluding the fundamental and full datasets, 2^k-2 partial prediction models are available. Moreover, a single full prediction model is shared by all fundamental combinations.

For every prediction model $M \in \{M_{x_1,x_2,\ldots x_k}, M_{(x_j,j\notin I\mid I\neq\emptyset \text{ and } I\neq\{1,2\ldots k\})}, M_{\Omega_1,\Omega_2,\ldots,\Omega_k}$ associated with a fundamental combination $(x_1, x_2, \ldots x_k)$, the prediction value $\tilde{y}^p_M$ can be generated for the p-th observation $\left(x^p_1, x^p_2, x^p_3, \ldots, x^p_k, x^p_{k+1}, \ldots, x^p_{k+m}, y^p\right) \in \Gamma_{x_1,x_2,\ldots x_k}$ using the prediction model M.

The training, validation, and testing $RMSE^M_{\Gamma^*_{x_1,x_2,\ldots x_k}}, * \in \{Tr, V, Ts$ of fundamental datasets $\Gamma^{Tr}_{\Omega_1,\Omega_2,\ldots\Omega_k}$, $\Gamma^{V}_{\Omega_1,\Omega_2,\ldots\Omega_k}$, $\Gamma^{Ts}_{\Omega_1,\Omega_2,\ldots\Omega_k}$ under model M can then be defined as

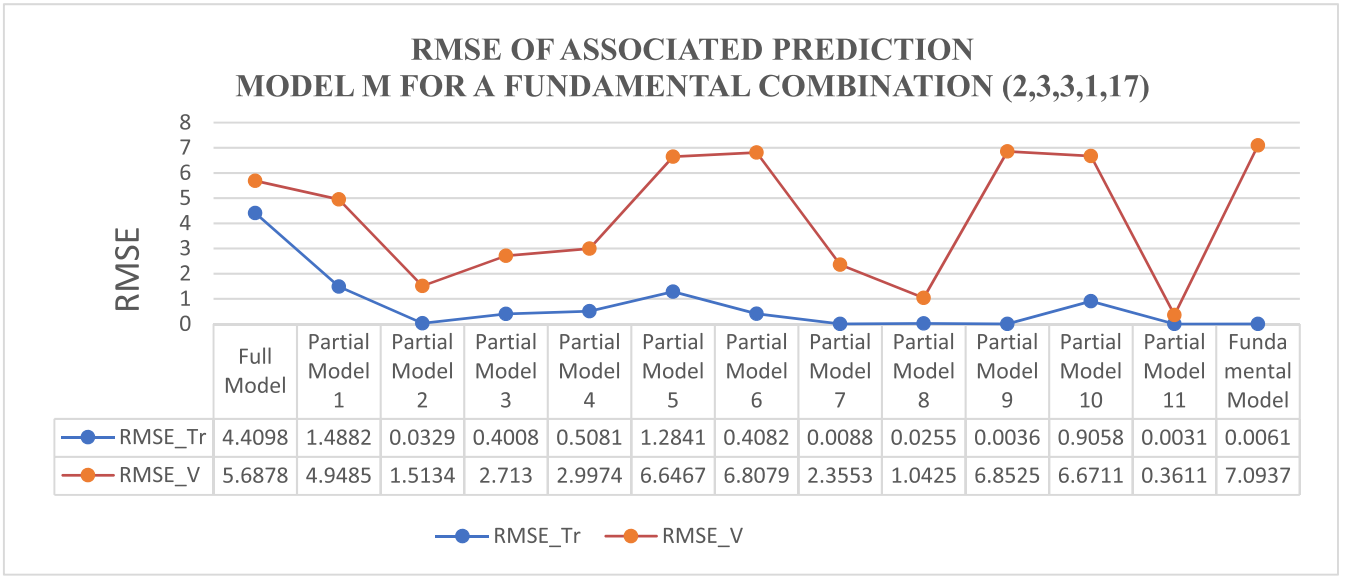
$$RMSE^M_{\Gamma^*_{x_1,x_2,\ldots x_k}} = \sqrt{\frac{\sum_{p\in\Gamma^*_{x_1,x_2,\ldots x_k}} (y^p - \tilde{y}^p_M)^2}{\text{number of observations in}\Gamma^*_{x_1,x_2,\ldots x_k}}},$$
$$* \in \{Tr, V, Ts\} \text{ and } M \in \{M_{x_1,x_2,\ldots x_k}, M_{(x_j,j\notin I\mid I\neq\emptyset \text{ and } I\neq\{1,2\ldots k\})}, M_{\Omega_1,\Omega_2,\ldots,\Omega_k},$$

where $RMSE^M_{\Gamma^{Ts}_{x_1,x_2,\ldots x_k}}$ is the RMSE of model M for the fundamental testing dataset $\Gamma^{Ts}_{x_1,x_2,\ldots x_k}$.

In practice, the numbers of distinct partial combination datasets and prediction models are less than $2^k - 2$ because many fundamental combinations are infeasible due to the incompatibility among categorical values (e.g., mismatch of the machine type and the materials/products). Infeasible fundamental combinations lead to empty fundamental combination datasets. When an empty dataset is in union with another fundamental/partial dataset, the new dataset remains unchanged, and the new models for the new dataset do not need to be trained. Thus, the number of distinct datasets and prediction models are less than what is theoretically allowed.

When multiple models are associated with each fundamental combination, different prediction models might exhibit various performance. For example, in the semiconductor manufacturing dataset used for the numerical analysis in this research, different fundamental/partial/full combination models are used to predict the production rates of a product. The training and validation RMSEs ($RMSE^M_{\Gamma^{Tr}_{x_1,x_2,\ldots x_k}}$ and $RMSE^M_{\Gamma^{V}_{x_1,x_2,\ldots x_k}}$) of the 13 distinct prediction models associated with a specific fundamental combination are plotted in Fig. 3, where the models are arranged according to the number of observations in the dataset used for model training from left to right (i.e., the full dataset contains all training data and is listed on the left, whereas the fundamental dataset is the smallest training dataset and the fundamental model is listed on the right).

In Fig. 3, one of the partial combination models possesses the lowest validation RMSE ($RMSE^M_{\Gamma^{V}_{x_1,x_2,\ldots x_k}}$), which indicates better prediction performance than those with higher validation RMSEs. Moreover, in the semiconductor dataset, the rightmost fundamental model demonstrates overfitting, and low training RMSE ($RMSE^{M_{x_1,x_2,\ldots x_k}}_{\Gamma^{Tr}_{x_1,x_2,\ldots x_k}}$) and high validation RMSE ($RMSE^{M_{x_1,x_2,\ldots x_k}}_{\Gamma^{V}_{x_1,x_2,\ldots x_k}}$) are observed. In addition, the full combination model $M_{\Omega_1,\Omega_2,\ldots,\Omega_k}$ does not perform satisfactorily because of the different interactions between the categorical and numerical attributes in the different datasets.



	Full Model	Partial Model 1	Partial Model 2	Partial Model 3	Partial Model 4	Partial Model 5	Partial Model 6	Partial Model 7	Partial Model 8	Partial Model 9	Partial Model 10	Partial Model 11	Fundamental Model
RMSE_Tr	4.4098	1.4882	0.0329	0.4008	0.5081	1.2841	0.4082	0.0088	0.0255	0.0036	0.9058	0.0031	0.0061
RMSE_V	5.6878	4.9485	1.5134	2.713	2.9974	6.6467	6.8079	2.3553	1.0425	6.8525	6.6711	0.3611	7.0937

FIGURE 3. $RMSE^{M}_{\Gamma^{*}_{x_1,x_2,\ldots x_k}}$ of the different models in the semiconductor manufacturing example.

Categorical attributes might have complex interaction with one another, as well as with other numerical attributes (Fig. 1 and Fig. 3); thus, a fundamental, partial, or full model can be appropriate or inappropriate for a specific fundamental combination of categorical attribute values. When some of the categorical attribute levels show similar influence on the response variable, combining datasets with such categorical attribute levels might create a large dataset that effectively estimates the model parameters. However, when the influence of the categorical attribute on the response variable vary, pooling the datasets together might mislead the model training processes and hamper the overall performance. In conclusion, estimating the response variable using the existing methods is difficult when complex interaction is possible in a mixed dataset.

VI. SELECTION OF THE PREDICTION MODELS

This section proposes a method that utilizes the training and validation datasets in selecting the appropriate prediction methods to overcome the weakness of the existing methods in mixed datasets with complex interaction. For each fundamental combination, the model selection method can select the appropriate prediction models among all models.

A. MODEL SELECTION INDEXES

First, we define the model selection indexes. For every record in the training and validation data in the fundamental training and validation datasets ($\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$ and $\Gamma^{V}_{x_1,x_2,\ldots x_k}$, respectively), each prediction model associated with the fundamental combination $(x_1, x_2, \ldots x_k)$ can be used to generate the prediction values. However, errors are present between the predicted and observed values. Using the prediction error of each dataset in each associated model, the RMSE and the 90th quantile (90QT) of errors for each possible pairs of an associated prediction model $M \in \{M_{x_1,x_2,\ldots x_k}, M_{(x_j,j\notin I|I\neq\emptyset \text{ and } I\neq\{1,2\ldots k\})}, M_{\Omega_1,\Omega_2,\ldots,\Omega_k}$ and a dataset $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$ or $\Gamma^{V}_{x_1,x_2,\ldots x_k}$ can be determined.

The performance measures of a prediction model to a dataset are listed in TABLE 4.

- Note 1: The 90 percent quantile (90 QT) of the prediction error is used for model selection because RMSE can amplify the influence or an outlier or glitch in the data collection processes. Thus, when outliers exist, 90 QT could serve as a robust model selection index.
- Note 2: The training dataset $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$, which contains 70% of the data, is a relatively stable measure of a model's prediction quality. However, using $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$ might fail to detect the overfitting of a model. Thus, the validation dataset $\Gamma^{V}_{x_1,x_2,\ldots x_k}$, which contains 15% of the data, is also used to measure the performance of a prediction model in a specific fundamental combination $(x_1, x_2, \ldots x_k)$.

B. ONE-STAGE AND TWO-STAGE MODEL SELECTION METHODS

The model fitness indexes (TABLE 4) for every distinct prediction model $\boldsymbol{M}$ associated with a fundamental combination are calculated and ranked from the best to the worst. The ranking is then used to define the one-stage and two-stage model selection methods.

1) ONE-STAGE MODEL SELECTION METHOD (A, B, N)

Let $A \in \{RMSE, 90QT$ be a performance measure, $B \in \{\Gamma^{Tr}_{x_1,x_2,\ldots x_k}, \Gamma^{V}_{x_1,x_2,\ldots x_k}\}$ be a dataset, and n be a positive integer within $1 \leq n \leq 2^k$. A one-stage model selection method identifies the best n models under performance index A using dataset B. For example, a one-stage (A, B, n) model selection method (RMSE, $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$, 3) ranks the top three models using RMSE and dataset $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$ (i.e., the top three models under

TABLE 4. Performance measure of a prediction model to a dataset.

Model Fitness Index	Definition
$RMSE^{M}_{\Gamma^{Tr}_{x_1,x_2,\ldots x_k}}$	The calculated RMSE of a prediction model M to the fundamental training dataset $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$.
$RMSE^{M}_{\Gamma^{V}_{x_1,x_2,\ldots x_k}}$	The calculated RMSE of a prediction model M to the fundamental validation dataset $\Gamma^{V}_{x_1,x_2,\ldots x_k}$.
$90QT^{M}_{\Gamma^{Tr}_{x_1,x_2,\ldots x_k}}$	The calculated 90QT of the absolute prediction error of a prediction model M to the fundamental training dataset $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$.
$90QT^{M}_{\Gamma^{V}_{x_1,x_2,\ldots x_k}}$	The calculated 90QT quantile of the absolute prediction error of a prediction model M to the fundamental validation dataset $\Gamma^{V}_{x_1,x_2,\ldots x_k}$.

$$The\ Model\ M \in \{M_{x_1,x_2,\ldots x_k}, M_{(x_j,j\notin I|I\neq\emptyset\ and\ I\neq\{1,2\ldots k\})}, M_{\Omega_1,\Omega_2,\ldots,\Omega_k}$$

$RMSE^{M}_{\Gamma^{Tr}_{x_1,x_2,\ldots x_k}}$ are selected). A_{B_n} is the set of the n models selected using the model selection method (A, B, n).

2) TWO-STAGE MODEL SELECTION METHODS (A, B, N, C, M)

Let $A \in \{RMSE, 90QT\}$ be a performance measure, $B \in \left\{\Gamma^{Tr}_{x_1,x_2,\ldots x_k}, \Gamma^{V}_{x_1,x_2,\ldots x_k}\right\}$ be a dataset and n be a positive integer within $1 \leq n \leq 2^k$, $C \in \left\{\Gamma^{Tr}_{x_1,x_2,\ldots x_k}, \Gamma^{V}_{x_1,x_2,\ldots x_k}\right\}$, where $C \neq B$, and m be another positive integer within $m < n \leq 2^k$.A two-stage model selection method can be defined as follows.

- Step 1: The best n models under performance index A are identified using dataset B (same as in the one-stage model selection method), and A_{B_n} is defined as the set of n models selected from the first stage.
- Step 2: The best m models in set A_{B_n} under performance index A are selected using dataset C, and $A_{B_nC_m}$ is defined as the set of models selected in this second stage.

For example, the two-stage model (RMSE, $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$, 10, $\Gamma^{V}_{x_1,x_2,\ldots x_k}$, 3) identifies the top 10 models using $RMSE^{M}_{\Gamma^{Tr}_{x_1,x_2,\ldots x_k}}$, which uses the dataset $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$ to calculate the RMSE. In Step 2, the top 3 models are selected among the 10 selected models using $\Gamma^{V}_{x_1,x_2,\ldots x_k}$ and RMSE.

The equality $C = B$ will degenerate a two-stage method into an equivalent one-stage method, hence, only the case $C \neq B$ is considered in the two-stage methods.

Let the set of selected prediction models from the one-stage model selection methods be A_{B_n}. The average prediction $\tilde{y}^p_{A_{B_n}}$ generated by A_{B_n} can then be expressed as

$$\tilde{y}^p_{A_{B_n}} = \frac{\sum_{M\in A_{B_n}} \tilde{y}^p_M}{n}.$$

Then, let the set of selected prediction models from the two-stage model be $A_{B_nC_m}$. The average prediction $\tilde{y}^p_{A_{B_nC_m}}$ generated by $A_{B_nC_m}$ can be expressed as

$$\tilde{y}^p_{A_{B_nC_m}} = \frac{\sum_{M\in A_{B_nC_m}} \tilde{y}^p_M}{m}.$$

C. PREDICTION QUALITY INDEXES FOR THE ONE-STAGE AND TWO-STAGE MODEL SELECTION METHODS

The corresponding testing RMSE, mean absolute percentage error (MAPE), 90QT, mean absolute error (MAE), mean arctangent absolute percentage error (MAAPE), and R-squared (R^2) of the model selection method # are determined, where # $\in\{A_{B_n}, A_{B_nC_m}\}$ is the selected selection method. For a fundamental combination $(x_1, x_2, \ldots x_k)$,the abovementioned performance indexes are defined as follows.

1)
$$RMSE^{\#}_{\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} = \sqrt{\frac{\sum_{p\in\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} (y^p - \tilde{y}^p_{\#})^2}{\text{number of observations in } \Gamma^{Ts}_{x_1,x_2,\ldots x_k}}}$$

2)
$$MAPE^{\#}_{\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} = \frac{\sum_{p\in\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} \left|\frac{y^p - \tilde{y}^p_{\#}}{y^p}\right|}{\text{number of observations in } \Gamma^{Ts}_{x_1,x_2,\ldots x_k}} \times 100\%$$

3)
$$90QT^{\#}_{\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} = \text{the 90\% quantile of prediction errors } |y^p - \tilde{y}^p_{\#}|$$

4)
$$MAE^{\#}_{\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} = \frac{\sum_{p\in\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} |y^p - \tilde{y}^p_{\#}|}{\text{number of observations in } \Gamma^{Ts}_{x_1,x_2,\ldots x_k}}$$

5)
$$MAAPE^{\#}_{\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} = \frac{\sum_{p\in\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} arctan\left|\frac{y^p - \tilde{y}^p_{\#}}{y^p}\right|}{\text{number of observations in } \Gamma^{Ts}_{x_1,x_2,\ldots x_k}}$$

6)
$$R^{2^{\#}}_{\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} = 1 - \frac{\sum_{p\in\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} (y^p - \tilde{y}^p_{\#})^2}{\sum_{p\in\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} (y^p - \bar{y})^2}$$

Then, the overall testing RMSE, MAPE, 90QT, MAE, and MAAPE are determined.

7)
$$RMSE^{\#}_{\Gamma^{Ts}} = \sqrt{\frac{\sum_{(x_1,x_2,\ldots\ldots x_k)} \sum_{p\in\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} (y^p - \tilde{y}^p_{\#})^2}{\text{number of observations in } \Gamma^{Ts}}}$$

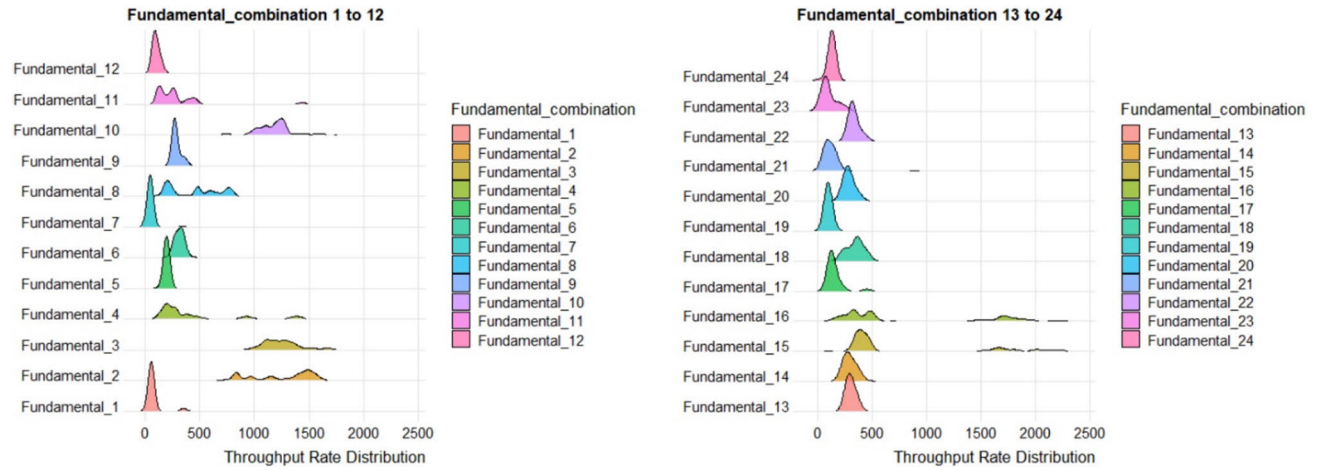


FIGURE 4. The throughput rate distribution within different fundamental datasets.

8)

$$MAPE^{\#}_{\Gamma^{Ts}} = \frac{\sum_{(x_1,x_2,\ldots..x_k)} \sum_{p\in\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} \left|\frac{y^p-\tilde{y}^p_{\#}}{y^p}\right|}{\text{number of observations in } \Gamma^{Ts}} \times 100\%$$

9)

$$90QT^{\#}_{\Gamma^{Ts}} = \text{ the } 90\% \text{ quantile of errors } |y^p - \tilde{y}^p_{\#}|$$

10)

$$MAE^{\#}_{\Gamma^{Ts}} = \frac{\sum_{(x_1,x_2,\ldots..x_k)} \sum_{p\in\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} \left|y^p - \tilde{y}^p_{\#}\right|}{\text{number of observations in } \Gamma^{Ts}}$$

11)

$$MAAPE^{\#}_{\Gamma^{Ts}} = \frac{\sum_{(x_1,x_2,\ldots..x_k)} \sum_{p\in\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} \arctan\left|\frac{y^p-\tilde{y}^p_{\#}}{y^p}\right|}{\text{number of observations in } \Gamma^{Ts}}$$

12)

$$R^{2\#}_{\Gamma^{Ts}} = 1 - \frac{\sum_{(x_1,x_2,\ldots..x_k)} \sum_{p\in\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} (y^p - \tilde{y}^p_{\#})^2}{\sum_{(x_1,x_2,\ldots..x_k)} \sum_{p\in\Gamma^{Ts}_{x_1,x_2,\ldots x_k}} (y^p - \bar{y})^2}$$

VII. NUMERICAL STUDY AND ANALYSIS

In this section, two datasets, Semiconductor Backend Production Rate [37] and Diamonds [38], are used to evaluate the proposed model selection prediction method.

A. SEMICONDUCTOR BACKEND PRODUCTION RATE DATASET

Semiconductor Backend Production Rate dataset [37] was collected between Oct. 2018 and Mar. 2019 from a world-leading semiconductor assembly and testing factory in Taiwan. This dataset includes five categorical attributes ($X_i, i = 1 \ldots 5$), 11 numerical attributes ($X_l, l = 6 \ldots 16$), and one response variable. The five categorical attributes represent the machine, product, material, package, and recipe types that are key factors affecting the production rates as recorded by the Manufacturing Execution System (MES) during the production. The 11 numerical attributes represent the geometric and physical characteristics of a semiconductor chip, such as the grinding thickness, number of wires, wire width and length, number of dies in a substrate, lead count, 2D die size, and the 3D package size. The response variable refers to the throughput rate of a specific machine–product combination during production. The five categorical variables X_1, X_2, X_3, X_4, X_5 have 2, 3, 4, 7 and 22 categorical attribute values, respectively (i.e., $\Omega_1 = \{1, 2\}$, $\Omega_2 = \{1, 2, 3\}$, $\Omega_3 = \{1, 2, 3, 4\}$, $\Omega_4 = \{1, 2, \ldots, 7\}$, and $\Omega_5 = \{1, 2, \ldots, 22\}$). In addition, the dataset contains 13,186 observations. Fig. 4 shows the throughput rate distribution within different fundamental datasets. Given the difference between fundamental datasets, estimating the throughput rate values by using existing prediction methods is difficult.

To the best of our knowledge, our research is among the first studies to adopt the model selection approach for prediction. The mixed dataset used might be helpful for other researchers who are interested in studying prediction problems with complex interaction among categorical attributes. Therefore, the dataset is provided as an electronic complementation of this paper to facilitate future research in this field.

The XGBoost package on R is used for the training of all fundamental/partial/full models. The following hyper-parameters are used for the training of all partial combination XGBoost models: general parameters {booster: gbtree}; booster parameters {eta: 0.03, gamma: 0, max_depth: 6, subsample: 0.85, colsample_bytree: 0.85}; and learning task parameters{objective:reg:linear, eval_metric:rmse}.

The proposed method using XGBoost package can internally handle missing values of numerical attributes. For categorical attribute with missing values, one-hot encoding allows the missing values to be treated as an additional level

TABLE 5. Testing performance in semiconductor dataset.

(A) Testing performance of different model selection methods

		One-stage model selection methods (A, B, n)								Two-stages model selection methods (A, B, n, C, m)							
model selection methods	A	RMSE	RMSE	RMSE	RMSE	90QT	90QT	90QT	90QT	RMSE	RMSE	RMSE	RMSE	90QT	90QT	90QT	90QT
	B	Tr	Tr	V	V	Tr	Tr	V	V	Tr	Tr	V	V	Tr	Tr	V	V
	n	1	3	1	3	1	3	1	3	10	10	10	10	10	10	10	10
	C	--	--	--	--	--	--	--	--	V	V	Tr	Tr	V	V	Tr	Tr
	M	--	--	--	--	--	--	--	--	1	3	1	3	1	3	1	3
model fitness indexes	$RMSE^{\#}_{\Gamma^{Ts}}$	74.11	68.57	70.03	71.39	74.01	68.77	72.78	66.66	72.30	70.51	75.35	71.81	72.16	67.10	75.64	68.07
	$MAPE^{\#}_{\Gamma^{Ts}}$	21.17	19.63	20.99	21.13	21.15	19.64	19.12	17.70	21.47	20.06	19.65	19.50	20.81	19.71	19.18	19.07
	$90QT^{\#}_{\Gamma^{Ts}}$	30.46	30.05	29.85	29.40	30.40	29.87	29.42	29.24	29.54	29.62	30.63	29.59	30.65	29.99	30.20	29.66
	$MAE^{\#}_{\Gamma^{Ts}}$	25.69	24.87	24.56	24.44	25.67	24.92	25.11	24.00	25.03	24.38	25.73	24.77	24.73	24.01	25.67	24.56
	$MAAPE^{\#}_{\Gamma^{Ts}}$	4.02	3.97	3.95	3.92	4.00	3.98	3.96	3.86	3.95	3.82	3.96	3.88	3.82	3.81	4.01	3.95
	$R^{2\#}_{\Gamma^{Ts}}$	0.9796	0.9825	0.9818	0.9811	0.9797	0.9824	0.9803	0.9835	0.9806	0.9815	0.9789	0.9809	0.9807	0.9833	0.9788	0.9828

(B) Testing performance against prediction methods from the literature

	Full Combination Regression Model	Fundamental Combination Regression Model	Full Combination XGBoost Model	Fundamental Combination XGBoost Model	Best One-Stage (90QT, $\Gamma^{V}_{x_1,x_2,\ldots x_k}$, 3) Method	Best Two-Stages (90QT, $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$, 10, $\Gamma^{V}_{x_1,x_2,\ldots x_k}$, 3) Method	Decision tree regression [13]	Polynomial Regression [13] (n=2)	Hierarchical Linear regression [16]
$RMSE^{\#}_{\Gamma^{Ts}}$	212.71	168.81	74.71	73.65	66.66	67.10	118.14	156.28	203.88
$MAPE^{\#}_{\Gamma^{Ts}}$	58.94	18.91	21.96	19.98	17.70	19.71	25.21	51.15	72.91
$90QT^{\#}_{\Gamma^{Ts}}$	204.34	64.08	30.35	30.20	29.24	29.99	28.64	101.85	118.17
$MAE^{\#}_{\Gamma^{Ts}}$	120.68	43.19	25.60	25.54	24.00	24.01	32.49	73.71	146.54
$MAAPE^{\#}_{\Gamma^{Ts}}$	25.80	10.38	4.31	4.11	3.86	3.81	6.71	19.1	28.94
$R^{2\#}_{\Gamma^{Ts}}$	0.8321	0.8942	0.9793	0.9799	0.9835	0.9833	0.9367	0.9303	0.8082

of the attribute. To ensure that reliable results are reported for different methods, cross-validation is performed by randomly selecting training, validation, and testing datasets five times on the semiconductor assembly and testing data.

Let $A \in \{RMSE, 90QT,\ B \in \{\Gamma^{Tr}_{x_1,x_2,\ldots x_k}, \Gamma^{V}_{x_1,x_2,\ldots x_k}\}$, and $n \in \{1, 3\}$ in the one-stage methods, and $A \in \{RMSE, 90QT,\ B \in \{\Gamma^{Tr}_{x_1,x_2,\ldots x_k}, \Gamma^{V}_{x_1,x_2,\ldots x_k}\}$, $n = 10$, $C \neq B \in \{\Gamma^{Tr}_{x_1,x_2,\ldots x_k}, \Gamma^{V}_{x_1,x_2,\ldots x_k}\}$, and $m \in \{1, 3\}$ in the two-stage model selection methods. Eight one-stage and eight two-stage methods can be generated. The testing results of these 16 model selection methods are reported in TABLE 5(A). The best performing one-stage model selection method is (90QT, $\Gamma^{V}_{x_1,x_2,\ldots x_k}$, 3), whereas the best performing two-stage model selection methods is (90QT, $\Gamma^{V}_{x_1,x_2,\ldots x_k}$, 10, $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$, 3). Both model selection methods selected three prediction models using the 90QT as the performance index. The findings revealed that this index is less sensitive to outliers than RMSE, and thus can serve as a robust index for model selection.

The best performing one- and two-stage model selection methods are compared with properly tuned full and fundamental combination XGBoost models (TABLE 5 (B)) to verify the superiority of the proposed model selection prediction method for mixed datasets. The hyper-parameters of the full and fundamental combination XGBoost models are tuned via grid search to ensure that fair comparison is achieved. We also compared other regression methods for mixed data sets, such as decision tree regression, polynomial regression, and hierarchical linear regression.

The results showed that the one-stage and two-stage model selection methods can effectively reduce prediction error, while all other models fail to perform satisfactorily in this numerical study because of the complex interaction among variables. Because a reliable fundamental combination model could not be trained using small data sets with less than 100 observations, the full combination model will be used by default for small data sets. Hence, TABLE 5 summarizes only results from the fundamental combinations with at least 100 observations. The RMSE improvement is ranging from 8.9% and 10.8% against the properly trained full or fundamental combination XGBoost models, which are used in the proposed method. The model selection approach also enhanced the prediction model transparency compared with the original XGBoost method, which is used for the training of all models. Although XGBoost is an ensemble decision tree method that provides better transparency than other ML methods [34], extracting explainable results is difficult when hundreds of different trees are present. Thus, when several models are selected, the similarity among these models can be used to help explain the ML results. For the illustration of model transparency, the 90QT and RMSE of the different models for the fundamental combination (2, 3, 3, 1, 17) of the semiconductor dataset are summarized (Fig. 5). The best performing model of the one-stage (90QT, $\Gamma^{V}_{x_1,x_2,\ldots x_k}$, 3) and two-stage selection methods (90QT, $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$, 10, $\Gamma^{V}_{x_1,x_2,\ldots x_k}$, 3) identified partial model 11($M_{(x_j, j\notin\{2\})}$), partial model 7($M_{(x_j, j\notin\{2,4\})}$) and partial model 2 ($M_{(x_j, j\notin\{2,3,4\})}$) for the prediction.

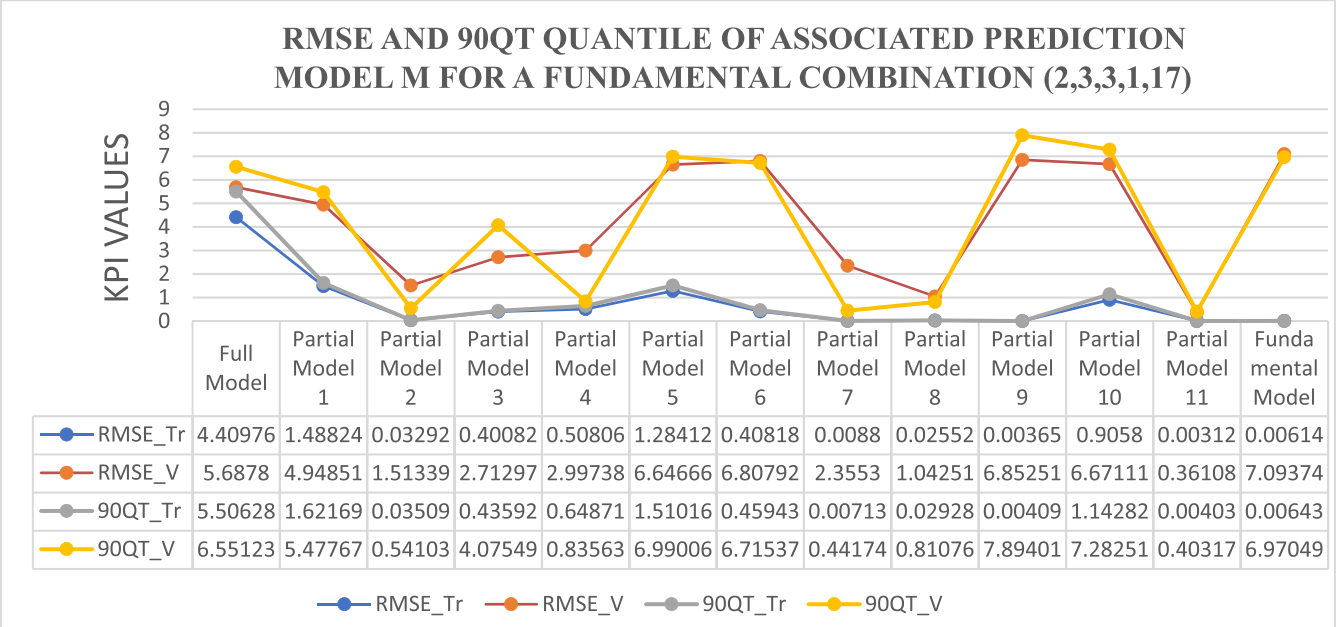


FIGURE 5. Enhancing the model's transparency using the model selection methods.

TABLE 6. Testing performance in Diamonds dataset.

(A)Testing performance of different model selection methods

		One-stage model selection methods (A, B, n)								Two-stages model selection methods (A, B, n, C, m)							
model selection methods	A	RMSE	RMSE	RMSE	RMSE	90QT	90QT	90QT	90QT	RMSE	RMSE	RMSE	RMSE	90QT	90QT	90QT	90QT
	B	Tr	Tr	V	V	Tr	Tr	V	V	Tr	Tr	V	V	Tr	Tr	V	V
	n	1	3	1	3	1	3	1	3	10	10	10	10	10	10	10	10
	C	--	--	--	--	--	--	--	--	V	V	Tr	Tr	V	V	Tr	Tr
	M	--	--	--	--	--	--	--	--	1	3	1	3	1	3	1	3
model fitness indexes	$RMSE^{\#}_{\Gamma^{Ts}}$	628.95	569.33	574.34	540.08	629.25	569.37	584.08	543.13	574.34	539.92	628.95	569.33	584.08	542.75	629.24	569.37
	$MAPE^{\#}_{\Gamma^{Ts}}$	7.7617	6.9212	6.9914	6.4767	7.7654	6.9213	7.0139	6.495	6.9914	6.4722	7.7617	6.9212	7.0139	6.5014	7.7656	6.9213
	$90QT^{\#}_{\Gamma^{Ts}}$	551.06	492.44	522.19	484.70	549.99	493.11	517.78	477.51	522.19	485.92	551.06	492.44	517.78	478.55	549.99	493.11
	$MAE^{\#}_{\Gamma^{Ts}}$	367.80	331.23	338.64	315.42	368.04	331.20	340.80	315.96	338.64	315.16	367.80	331.23	340.80	316.25	368.05	331.20
	$MAAPE^{\#}_{\Gamma^{Ts}}$	7.6731	6.8546	6.9261	6.4215	7.6767	6.8547	6.9455	6.4379	6.9261	6.4169	7.6731	6.8546	6.9455	6.4442	7.6769	6.8547
	$R^{2\#}_{\Gamma^{Ts}}$	0.9678	0.9736	0.9731	0.9762	0.9677	0.9736	0.9722	0.9760	0.9731	0.9762	0.9678	0.9736	0.9722	0.976	0.9677	0.9736

(B) Testing performance against prediction methods from the literature

	Full Combination Regression Model	Fundamental Combination Regression Model	Full Combination XGBoost Model	Fundamental Combination XGBoost Model	Best One-Stage (RMSE, $\Gamma^{V}_{x_1,x_2,\ldots x_k}$, 3) Method	Best Two-Stages (90QT, $\Gamma^{Tr}_{x_1,x_2,\ldots x_k}$, 10, $\Gamma^{V}_{x_1,x_2,\ldots x_k}$, 3) Method	Decision tree regression [13]	Polynomial Regression [13] (n=2)	Hierarchical Linear regression [16]
$RMSE^{\#}_{\Gamma^{Ts}}$	1117.17	1274.77	552.43	629.57	540.08	542.75	777.55	1018.73	1148.78
$MAPE^{\#}_{\Gamma^{Ts}}$	27.70	13.47	6.67	7.76	6.47	6.50	9.16	18.35	44.47
$90QT^{\#}_{\Gamma^{Ts}}$	1206.23	562.52	502.55	549.09	484.70	478.55	499.00	722.59	803.95
$MAE^{\#}_{\Gamma^{Ts}}$	747.04	453.16	325.87	367.72	315.42	316.25	373.42	549.50	1267.75
$MAAPE^{\#}_{\Gamma^{Ts}}$	21.84	11.84	6.61	7.44	6.42	6.44	8.91	16.88	31.72
$R^{2\#}_{\Gamma^{Ts}}$	0.8949	0.8632	0.9751	0.9677	0.9762	0.9760	0.9612	0.9334	0.9161

Among the three selected models, partial model 11 ($M_{(x_j, j\notin\{2\})}$) pools together the fundamental datasets with different categorical attribute (X_2) values. Therefore, different X_2 attributes exert similar influences on the response variable, and pooling these training datasets improves the prediction accuracy (i.e., the categorical attribute combinations $(x_1, x_2, \ldots x_5) = (2, \Omega_2, 3, 1, 17)$ influence the response variable in the same way). Similarly, for partial

models 7 and 2, the categorical attribute combinations $(2, \Omega_2, 3, \Omega_4, 17)$ and $(2, \Omega_2, \Omega_3, \Omega_4, 17)$ yield similar prediction models. In other words, different X_2, X_3, and X_4 combinations will not affect the prediction model behavior when $x_1 = 2$ and $x_5 = 17$. Moreover, the model selection results suggest that the categorical variables' X_1 and X_5 values interact or exert different influences on the response variable. When a partial combination model is selected, the similarity between fundamental combination data sets is revealed. As shown in Fig. 4, fundamental combinations 1 and 7 have a similar response variable distribution. The fundamental data sets are combined to create a large partial combination data set that would enhance the prediction accuracy and suggest the hidden correlation between those categorical attributes. Enhancing model transparency using partial combination of categorical attributes is among the first in literature.

B. DIAMONDS DATASET

To validate the proposed method further, we compare with the existing approaches using the open data set, Diamonds from ggplot2 [38]. The data set contains 53,940 observations, and further details can be found in [39]. This data set includes 3 categorical attributes $(X_i, i = 1 \ldots 3)$, which reflect the quality of the cut of the diamonds, diamonds' color, and a measurement of the extent of clarity of the diamond, and six numerical attributes $(X_l, l = 4 \ldots 9)$. The response variable refers to the price of diamonds in US dollars. In the Diamonds data set, the categorical variables X_1, X_2, X_3, have 5, 7, and 8 categorical attribute values, respectively (i.e., $\Omega_1 = \{1, 2, 3, 4, 5\}$, $\Omega_2 = \{1, 2, 3, 4, 5, 6, 7\}$ and $\Omega_3 = \{1, 2, 3, 4, 5, 6, 7, 8\}$). The proposed model selection method is compared with the results in [40] and several other methods from the literature. The results are summarized in TABLE 6. According to the model fitness indexes shown in TABLE 6, similar performance improvement is observed in this open mixed data set. The performance improvement in RMSE ranges from 1.8% and 14.2% against the properly trained full or fundamental combination XGBoost model, which is used in the proposed method and the performance improvement in RMSE is at least 30.2% against other methods from the literature.

VIII. CONCLUSIONS AND FUTURE WORK

In this study, a novel model selection method is proposed to improve the prediction performance for mixed datasets with complex interactions and the transparency of the prediction method. The proposed model selection prediction method is compatible with any existing regression or ML prediction method. Multiple prediction models can be generated under different categorical attribution combinations by partitioning a dataset into subsets with different categorical attribution combinations. One-stage and two-stage model selection methods are applied to the training and validation datasets to select the appropriate models. Results demonstrated the potential of the proposed model selection prediction methods in the mixed dataset. The cross-validation test results indicated a 10% improvement in the prediction accuracy with respect to the properly tuned XGBoost models. Moreover, compared with other methods from the literature, at least 30% reduction in RMSE is observed when the proposed methods are applied. In the future, the relationship of the model selection parameters (A, B, n, C, m) to the characteristics of the datasets will be explored. The potential of the model selection method can be further enhanced by properly tuning such parameters. Moreover, different statistics or ML methods will also be combined in the model training stages to further improve the overall prediction accuracy.

ACKNOWLEDGMENT

This study was conducted in part under the research projects, "Advanced Artificial Intelligence Technologies and Industry Applications (3/4)" with Grant No. 109-EC-17-A-21-1516 and "III Innovative and Prospective Technologies Project (1/1)" with Grant No. 109-EC-17-A-24-0461, which are subsidized by the Ministry of Economic Affairs in Taiwan. The work of the corresponding author was supported in part by the Ministry of Science and Technology, Taiwan under Grant No. MOST107-2628-E-002-006-MY3.

REFERENCES

[1] N. Geng, Z. Jiang, and F. Chen, "Stochastic programming based capacity planning for semiconductor wafer fab with uncertain demand and capacity," *Eur. J. Oper. Res.*, vol. 198, no. 3, pp. 899–908, Nov. 2009.

[2] P. Backus, M. Janakiram, S. Mowzoon, G. C. Runger, and A. Bhargava, "Factory cycle-time prediction with a data-mining approach," *IEEE Trans. Semicond. Manuf.*, vol. 19, no. 2, pp. 252–258, May 2006.

[3] D. Y. Sha, R. L. Storch, and C.-H. Liu, "Development of a regression-based method with case-based tuning to solve the due date assignment problem," *Int. J. Prod. Res.*, vol. 45, no. 1, pp. 65–82, Jan. 2007.

[4] L. Lingitz, V. Gallina, F. Ansari, D. Gyulai, A. Pfeiffer, W. Sihn, and L. Monostori, "Lead time prediction using machine learning algorithms: A case study by a semiconductor manufacturer," *Procedia CIRP*, vol. 72, pp. 1051–1056, 2018.

[5] B. Pavlyshenko, "Machine learning, linear and Bayesian models for logistic regression in failure detection problems," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 2046–2050.

[6] A. Mangal and N. Kumar, "Using big data to enhance the bosch production line performance: A Kaggle challenge," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 2029–2035.

[7] P. Su, Y. Liu, and X. Song, "Research on intrusion detection method based on improved smote and XGBoost," in *Proc. 8th Int. Conf. Commun. Netw. Secur. (ICCNS)*, 2018, pp. 37–41.

[8] P. M. Lerman, "Fitting segmented regression models by grid search," *Appl. Statist.*, vol. 29, no. 1, pp. 77–84, 1980.

[9] B. Ari and H. A. Güvenir, "Clustered linear regression," *Knowl.-Based Syst.*, vol. 15, no. 3, pp. 169–175, Mar. 2002.

[10] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[11] B. Zhang, "Regression clustering," in *Proc. 3rd IEEE Int. Conf. Data Mining*, Nov. 2003, pp. 451–458.

[12] Y. W. Park, Y. Jiang, D. Klabjan, and L. Williams, "Algorithms for generalized clusterwise linear regression," *INFORMS J. Comput.*, vol. 29, no. 2, pp. 301–317, May 2017.

[13] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, NY, USA: Springer, 2013, doi: 10.1007/978-1-4614-7138-7.

[14] S. Salvador and P. Chan, "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms," in *Proc. 16th IEEE Int. Conf. Tools with Artif. Intell.*, Nov. 2004, pp. 576–584.

[15] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. Kdd*, Aug. 1996, vol. 96, no. 34, pp. 226–231.
[16] J. V. Petrocelli, "Hierarchical multiple regression in counseling research: Common problems and possible remedies," *Meas. Eval. Counseling Develop.*, vol. 36, no. 1, pp. 9–22, Apr. 2003.
[17] H. Woltman, A. Feldstain, J. C. MacKay, and M. Rocchi, "An introduction to hierarchical linear modeling," *Tuts. Quant. Methods Psychol.*, vol. 8, no. 1, pp. 52–69, 2012.
[18] H. Späth, "Algorithm 39 clusterwise linear regression," *Computing*, vol. 22, no. 4, pp. 367–373, 1979.
[19] B. Chizi and O. Maimon, "Dimension reduction and feature selection," in *Data Mining and Knowledge Discovery Handbook*. Boston, MA, USA: Springer, 2009, pp. 83–100, doi: 10.1007/978-0-387-09823-4_5.
[20] C. Quan, D. Wan, B. Zhang, and F. Ren, "Reduce the dimensions of emotional features by principal component analysis for speech emotion recognition," in *Proc. IEEE/SICE Int. Symp. Syst. Integr.*, Dec. 2013, pp. 222–226.
[21] J. M. Cadenas, M. C. Garrido, and R. Martínez, "Feature subset selection Filter–Wrapper based on low quality data," *Expert Syst. Appl.*, vol. 40, no. 16, pp. 6241–6252, Nov. 2013.
[22] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben, "Machine learning in manufacturing: Advantages, challenges, and applications," *Prod. Manuf. Res.*, vol. 4, no. 1, pp. 23–45, Jan. 2016.
[23] D. Wu, S. Liu, L. Zhang, J. Terpenny, R. X. Gao, T. Kurfess, and J. A. Guzzo, "A fog computing-based framework for process monitoring and prognosis in cyber-manufacturing," *J. Manuf. Syst.*, vol. 43, pp. 25–34, Apr. 2017.
[24] D. Wu, C. Jennings, J. Terpenny, R. X. Gao, and S. Kumara, "A comparative study on machine learning algorithms for smart manufacturing: Tool wear prediction using random forests," *J. Manuf. Sci. Eng.*, vol. 139, no. 7, Jul. 2017, Art. no. 071018.
[25] D. Moldovan, T. Cioara, I. Anghel, and I. Salomie, "Machine learning for sensor-based manufacturing processes," in *Proc. 13th IEEE Int. Conf. Intell. Comput. Commun. Process. (ICCP)*, Sep. 2017, pp. 147–154.
[26] D. Stanisavljevic and M. Spitzer, "A review of related work on machine learning in semiconductor manufacturing and assembly lines," in *Proc. SAMI iKNOW*, 2016, pp. 1–6.
[27] P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, "A detailed investigation and analysis of using machine learning techniques for intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 686–728, 1st Quart., 2019.
[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.
[29] H. Zheng, J. Yuan, and L. Chen, "Short-term load forecasting using EMD-LSTM neural networks with a XGBoost algorithm for feature importance evaluation," *Energies*, vol. 10, no. 8, p. 1168, Aug. 2017.
[30] J. Zhong, Y. Sun, W. Peng, M. Xie, J. Yang, and X. Tang, "XGBFEMF: An XGBoost-based framework for essential protein prediction," *IEEE Trans. Nanobiosci.*, vol. 17, no. 3, pp. 243–250, Jul. 2018.
[31] L. Torlay, M. Perrone-Bertolotti, E. Thomas, and M. Baciu, "Machine learning–XGBoost analysis of language networks to classify patients with epilepsy," *Brain Inform.*, vol. 4, no. 3, p. 159, 2017.
[32] L. D. Raedt, K. Kersting, S. Natarajan, and D. Poole, "Statistical relational artificial intelligence: Logic, probability, and computation," *Synth. Lectures Artif. Intell. Mach. Learn.*, vol. 10, no. 2, pp. 1–189, Mar. 2016.
[33] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1675–1684.
[34] K. Broelemann and G. Kasneci, "A gradient-based split criterion for highly accurate and transparent model trees," 2018, *arXiv:1809.09703*. [Online]. Available: http://arxiv.org/abs/1809.09703
[35] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surveys*, vol. 51, no. 5, pp. 1–42, 2018.
[36] S. Putatunda and K. Rama, "A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost," in *Proc. Int. Conf. Signal Process. Mach. Learn. (SPML)*, 2018, pp. 6–10.
[37] C. J. Yu. *Semiconductor Backend Production Rate Dataset With Partial Combination Models*. Accessed: May 23, 2020. [Online]. Available: https://github.com/fishyu-tw/Semiconductor-Backend-Production-Rate-Dataset-with-Partial-Combination-Models
[38] *Prices of 50,000 Round Cut Diamonds From Ggplot2 Package*. Accessed: May 23, 2020. [Online]. Available: https://vincentarelbundock.github.io/Rdatasets/datasets.html
[39] Ggplot2 Documentation. *Diamonds: Prices of Over 50,000 Round Cut Diamonds*. Accessed: May 23, 2020. [Online]. Available: https://rdrr.io/cran/ggplot2/man/diamonds.html
[40] S. Panwala. *Regression-Based Machine Learning Approaches for Diamond Price Prediction*. Accessed: May 23, 2020. [Online]. Available: https://medium.com/@sp7091/regression-approaches-to-predict-diamond-price-258478a485c9

YI-HSIN WU (Associate Member, IEEE) received the B.S. degree, in 2007. She is currently pursuing the Ph.D. degree in electrical engineering with National Taiwan University, Taipei, Taiwan. Since 2012, she has been working at the Institute for Information Industry, Taipei, where she is a Principal Engineer leading the development of smart manufacturing solutions for the semiconductor packaging and testing industries. Her research interests include streaming analytics, parallel and distributed computing, embedded systems, algorithms, and artificial intelligence. She and her team were a recipient of the 56th Annual Research and Development 100 Awards Winner, Orlando, USA, in November 2018.

YU-HSIN CHANG received the bachelor's degree from the Department of Statistics, National Cheng Kung University, Tainan, Taiwan, in 2018. She is currently pursuing the master's degree with the Institute of Industrial Engineering, National Taiwan University, Taipei, Taiwan. Her research interests include about statistical data analysis, machine learning, and optimization.

YIN-JING TIEN received the B.S. degree in applied mathematics and the M.S. degree in statistics from National Chiao Tung University, Taiwan, in 1997 and 1999, respectively, and the Ph.D. degree in statistics from National Central University, Taiwan, in 2010. From 2010 to 2014, he was a Postdoctoral Researcher with the Institute of Statistical Science, Academia Sinica, Taiwan. He is currently a Principal Engineer of the Digital Transformation Institute with the Institute for Information Industry, Taiwan. His current research interests include statistical modeling, machine learning, artificial intelligent, and data mining.

CHENG-JUEI YU received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 2006 and 2011, respectively. Since 2011, he has been working at the Institute for Information Industry, Taipei, where he is currently a Section Manager leading the development of streaming analytics solutions for industrial applications. His research interests include streaming analytics, distributed computing, and parallel algorithms. He was a recipient of the Outstanding Young Engineer from the Chinese Institute of Engineers, Taipei, in 2018.

SHENG-DE WANG (Member, IEEE) received the B.S. degree from National Tsing Hua University, Hsinchu, Taiwan, in 1980, and the M.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1982 and 1986, respectively. Since 1986, he has been on the faculty of the Department of Electrical Engineering, National Taiwan University, where he is currently a Professor. From 1995 to 2001, he was the Director of the Computer and Information Network Center, Computer Operating Group, National Taiwan University. He was a Visiting Scholar with the Department of Electrical Engineering, University of Washington, Seattle, from 1998 to 1999. From 2001 to 2003, he was the Department Chair of the Department of Electrical Engineering, National Chi Nan University, Puli, Taiwan. His research interests include embedded systems, Internet computing and security, and intelligent systems.

CHENG-HUNG WU (Member, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1998, and the M.S. and Ph.D. degrees in industrial and operations engineering from the University of Michigan, Ann Arbor, in 2004 and 2006, respectively. Since 2007, he has been on the faculty of the Institute of Industrial Engineering, National Taiwan University, where he is currently an Associate Professor. His research interests include decisions under uncertainties, information and decision support systems, theoretical work in Markov decision processes, stochastic programming, and data science.