Feature Selection to Improve Performance of Yield Prediction in Hard Disk Drive Manufacturing

Anusara Hirunyawanakul[1], Nuntawut Kaoungku[2], Nittaya Kerdprasop[2], and Kittisak Kerdprasop[2]
[1] School of Computer Engineering, Suranaree University of Technology, Thailand
[2] Data and Knowledge Engineering Research Unit, Suranaree University of Technology, Thailand
Email: anusara.hi@gmail.com; {nuntawut; nittaya; kerdpras}@sut.ac.th

Abstract—**Hard Disk Drive (HDD) manufacturing is one real-world application area that machine learning has been extensively adopted for problem solving. However, most problem solving activities in HDD industry tackle on failure root-cause analysis task. Machine learning is rarely applied in a task of yield prediction. This research presents the application of machine learning and statistical techniques to select appropriate features to be used in yield prediction for the HDD manufacturing process. The seven well-known algorithms are used in the feature selection step. These algorithms are decision tree (C5 and CART), Support Vector Machine (SVM), stepwise regression, Genetic Algorithm (GA), chi-square and information gain. The two prominent learning algorithms, Multiple Linear Regression (MLR) and Artificial Neural Networks (ANN), are used in the yield prediction modeling step. Yield prediction performance has been assessed based on the two evaluation metrics: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Yield prediction with MLR shows higher accuracy than yield estimation traditionally performed by human engineers. Resulting to conclusion that the proposed novel learning steps can help HDD process engineers to predict yield with the better performance, especially on applying GA as feature selection tool, the MAE is reduced from 0.014 (yield estimated by human engineer) to 0.0059 (yield predicted by MLR). That means error reduction is about 60%.**

Index Terms—**Artificial neural network, feature selection, genetic algorithm, hard disk drive, multiple linear regression, yield prediction**

I. Introduction

Hard Disk Drive (HDD) is still be the most important data storage device and more preferred in the current era of big data than Solid State Drives (SSD). This is due to the two key factors, reliability of data retention and cost per terabyte that makes HDD clearly better than SSD. Even though SSD and flash memory are commonly used in personal digital devices, the enormous amount of data is stored in the server farms driven by ensemble of a lot of HDDs [1]-[2].

The production process of HDDs consists of many steps. Every single unit of a complete HDD is assembled from several components and required significant periods of time to check its quality. In some product series of HDDs with high capacity (such as 14TB or 12TB), the quality control may consume time of test process for over 3 months. The HDDs that pass the test process are called the "passed units," whereas the rejected HDDs are called the "failed units." It is certain that manufacturing factory prefers to produce "passed units" as much as possible. The metric to measure efficiency of production is "passed units" per "input units." This metric is commonly referred to as "yield" [3]-[5].

HDD manufacturers have to make strategic planning such as production manufacturing line planning, material usage planning, tester and machinery capacity planning, and shipment planning accurately. All of these planning activities involve the action of "yield prediction". For example, the precise yield prediction leads to suitable stocking material and optimal workforce and tester capacity plan. In current HDD manufacturing, typical method for yield prediction practice is based on personal experiences of process engineers. Even though many machine learning techniques are applied to assist the HDD manufacturing, they are focused on only failure analysis task. As far as we know, there are no application to the yield prediction task.

The most difficult portion of yield prediction task is the excessive amount of attributes obtained from several steps and components along the assembly and test process of HDD production manufacturing line. At least over 100 attributes are generated for a unit of HDD. It is likely impossible to compute and consider all attributes in the modeling step of yield prediction task [6]-[11].

In this paper, we thus propose data preparation and feature selection methods with the main focus to improve modeling performance on predicting yield. Many machine learning algorithms and techniques are used in this paper including support vectors machine (SVM), classification and regression tree (CART), C5, feature selection by considering chi-square and information gain.

Manuscript received December 30, 2019; revised January 15, 2020; accepted March 16, 2020.

Corresponding author: Anusara Hirunyawanakul (email: anusara.hi@gmail.com).

This research work has been supported by grants from the National Research Council of Thailand (NRCT). Data and Knowledge Engineering Research Unit has been fully supported by a research grant from Suranaree University of Technology.

doi: 10.18178/ijeetc.9.6.420-428

Our intuitive idea is to adopt these algorithms and techniques to select only important attributes to use in yield prediction step and expect to see the better performance of yield prediction.

In the part of yield prediction, we use two algorithms: Multiple Linear Regression (MLR) and Artificial and Neural Networks (ANN). Dataset used in our experiments contains around 1,000,000 records of HDD units that had been tested within one year. We group these million records into 1,000 rows of 10,000 records per group and use this new dataset for yield prediction step.

The next section of this paper presents briefly the background information regarding HDD components, the HDD production steps, and yield calculation in HDD manufacturing. Feature selection algorithms, yield prediction algorithm and evaluation method are also described in this section as well. In Section III, we explain material and methods that we use to develop the feature selection models. Section IV explains our experimental setting and results. The conclusion of this paper is in section V. Finally, in section VI, we provide the suggestion and recommendation on applying our idea.

II. Background and Theory

A. Hard Disk Drive (HDD)

HDD is a digital data storage device which records data on the durable platter (or hard disk) by magnetic recording technology. HDD is non-volatile storage device, which means HDD is able to store data even if power is off [12]-[14]. HDD consists of numerous important hardware components working together in a synchronized manner. Synchronization speed can affect the read-write performance. The fundamental components of HDD are shown in Fig. 1 and can be described as follows:

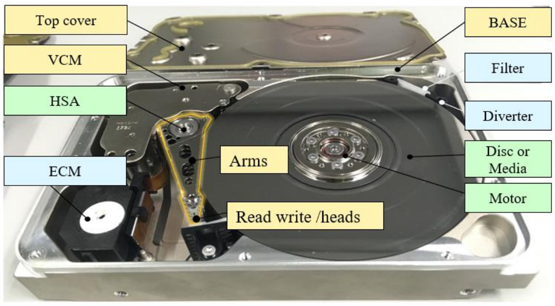


Fig. 1. Key components of a hard disk drive.

1) HSA (head stack assembly) [15] is the assembled part of reader/writer heads and base plate of the head components. HSA moves synchronously with rotation motor speed which drives the disk.

2) Media, disk or platter [16] is the key component for storing data. The digital data are written down from magnetic head to magnetic layer of disk. The substrate of disk component must be casted from durable and very smooth material such as aluminum or glass.

3) Motor base assembled (MBA) [17] is a component that consists of motor and motor hub plate. The key function of motor is to rotate the disk with consistency speed according to the read-write speed of each product revision. Motor hub plate is the strongest component of HDD because its function is to protect other components from external force.

4) Voice coil motor (VCM) [18] is consisted of two pieces of permanent magnet. This component works together with HSA to move HSA to the desired area for reading/writing data. This component function is working based on principle of magnetic field.

5) Printed circuit board assembled (PCBA) [19] is the controller of HDD. It is composed of several circuit wires, capacitance and microcontroller chip. Key function of this component is to communicate to computer or tester slot.

There exist other components that are also important such as environment control module (ECM) for controlling environment in the internal closed humidity and air flow HDD, Recirculate Filter for filtering out small particles from contaminating the HDD, Top-cover for sealing and enclosing the HDD. There are many other components with the main function to deliver the most reliable data storage as much as possible.

All of these components are assembled in the clean room called class-100, which is control count of particle size 500 nm to be lower than 100 counts. After completing the assembled process, the HDD is input into the "test process" to validate that the completed HDD is working properly in terms of mechanical and electrical properties, data storage, read-write performance, degradation and cosmetic outlook. In some product series that have high capacity in a unit, a complete flow of test process can be longer than 40 days with over 10 operation steps of test. The very long testing time is because the HDD maker must ensure that each of the many components work properly not only on a function of itself, but also operates synchronously with other components. The HDD that passes the test process is called the "pass unit," while the one rejected by test process is called the "fail unit."

B. Yield Definition and Its Calculation

Definition of "yield" in HDD manufacturing is simple and straightforward. It is a ratio between "pass unit" and "input unit" in the particular process or step. The calculation of yield is described as

$$\text{Yield}=\frac{\text{Quantity of pass units}}{\text{Quantity of input units}} \qquad (1)$$

Every single HDD is composed of many components that are assembled through many steps. Therefore, there are many attributes generated from each steps of the assembly process. These attributes are important factors in yield calculation. Thus, prior to the modeling process for yield prediction, one necessary task is to find and select only the important attributes for yield prediction.

C. Feature Selection

Feature selection is an important data preparation step before employing any machine learning algorithms or statistical analysis methods [20]-[22]. The objective of feature selection is to reduce the dimension of data to a manageable and computational size. Fundamental idea of feature selection is to find only the most powerful and discriminative attributes from many existing attributes or features. Feature selection techniques can be categorized into two major types: filter and wrapper.

Filter method performs attribute selection as a preprocessing step independent from a modeling step. Attributes are evaluated to select only the ones expected to have the most impact or importance on predicting the target attribute. Then those attributes will be prioritized in descending order. The threshold will be determined. If any of the important features do not reach the specified threshold, they will be discarded because of the assumption that they are not important enough. This method can be done in both univariate filter method and multivariate filter method. The most popular criteria for feature selection are chi-square and information gain. The advantages of filter method are simple calculation steps, fast computation, and the avoidance of overfitting problem.

Wrapper method, on the contrary, is tightly couple to the modeling step. The principle of this method is to take multiple features into consideration in the format of a set of features. Then, the modeling algorithm tries to find the feature set showing the most important association to the output feature. After the best feature set has been found, the learning algorithm will use this set in the subsequent step. There are 2 subtypes of wrapper method:

1) Forward stepwise is to continuously add more features one by one until the modeling algorithm can get the best set of attributes.

2) Backward stepwise is to put all the features in the set and then continuously take feature out of the set one by one until the best set is achieved.

The most popular methods for of wrapper feature selection method are stepwise regression and genetic algorithm.

The advantages of the wrapper method are simplicity and high efficiency. The disadvantage is that this method takes more time than the filter method and may cause overfitting problem.

D. Algorithm for Feature Selection: GA

Genetic Algorithm (GA) [23]-[25] is a technique for finding an optimal solution or approximate answers to a problem based on the theory of evolution from biology and natural selection. That is, the most suitable organisms can survive. GA consists of 5 main functions:

1) Chromosome encoding is taking the features of possible answers into a form of chromosome.

2) Initial population is to define the number of populations that we would like to create, usually done by randomized. Then, chromosomes are randomly generated by that amount.

3) Fitness function is to identify the function to be used for determining which chromosomes should go to next round. The criteria for justification will be different for each problem.

4) Genetic operator (selection, crossover and mutation) is a method of adjusting the structure of the chromosomes for the next model.

5) Termination is the function to define the point that we are satisfied, such as the best fitness score is achieved or the score is steady for many generations consecutively.

E. Algorithm for Feature Selection: Decision Tree

Decision Tree is a widely known algorithm for data classification algorithm. The concept of decision tree is finding the pattern of the attribute that needs to be classified. The structure of the data classification model is in the hierarchy [26], [27] and represented as a tree. The tree is consisted of nodes and branches. Nodes can be divided into 3 types as root node, internal node and leaf node as shown in Fig. 2.

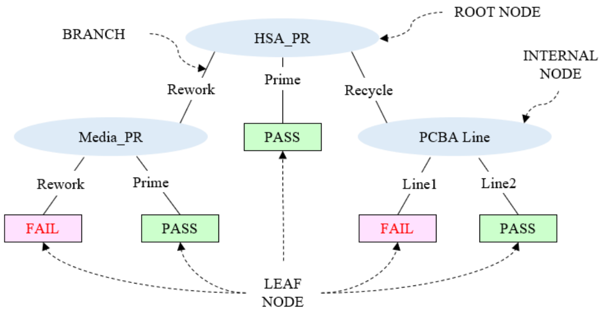


Fig. 2. Example structure of decision tree in HDD manufacturing.

Root nodes and internal nodes are features that the learning algorithm selects for being decision criteria for splitting data of mixed classes to be data subsets with more purity of class mixture. The root node is the initial feature chosen by the algorithm to create a tree. The next step is creating the branches of the root node. The number of branches will equal to all possible values of the features selected as the root node. If any child node contains data having the same class, then that node becomes a leaf node. Conversely, if data in the child node are of mixed classes, the tree is growing by repeating the splitting process until all leaf nodes are of homogeneous class or until some stopping criterion has been met. In this work, we apply the decision tree as one of our feature selection methods. There are many criteria for splitting nodes in a decision tree as shown in Table I.

TABLE I: EXAMPLE OF ALGORITHMS AND CRITERIA FOR DECISION TREE MODELING

Criteria for construct decision tree	Algorithm name
Information Gain	ID3, C4.5, C5.0
Gini Index	CART
Chi-Square	CHAID
Variance Reduction	CART

F. Algorithm for Yield Prediction: MLR

Regarding to literature review on yield prediction, the one prominent algorithm is linear regression because of its simplicity and predictive performance. Linear regression [28]-[30] is the statistical method seeking for quantitative correlation among two or more variables. One variable has to be defined as a target of analysis; this variable is called dependent variable, denoted with a common symbol Y. The other variables are used for predicting the value of a target variable; these variables are called independent variables, denoted as X_i, when i is 1 to k for the case that there exist k independent variables. The modeling of linear regression is based on the calculation defined as

$$\hat{Y} = a + bX \tag{2}$$

where $\hat{Y}$ is the dependent variable (or target variable for prediction), X is the independent variable, a is the constant of regression (or cutting point on Y axis), and b is the slope of a line (or regression coefficient of X).

In (2) we assume there is only one independent variable. The computation processing of this linear regression is to find the best coefficient of variable X and some constant value to predict the value of variable Y with least error. The relation of variable X and Y can be plotted with linear graph. This computation is also called a simple linear regression analysis.

In case of multiple independent variables, the modeling will be called multiple linear regression (MLR) analysis. The computation of MLR can be done by

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k \tag{3}$$

where $\hat{Y}$ is the dependent variable, (X_1, X_2, ···, X_k) is a set of k independent variables, b_0 is a constant of regression (or cutting point on Y axis), and (b_1, b_2, ···, b_k) is a set of line's slopes (or regression coefficients of the k independent variables).

G. Algorithm for Yield Prediction: ANN

Besides linear regression analysis, machine learning is another new technology being adopted as an interesting alternative method for yield forecasting. The most popular machine learning technique used in yield prediction is artificial neural network (ANN). Popularity is due to its outstanding performance. ANN is machine learning algorithm that is inspired by the biological neural networks of brains [31]-[33]. There are plenty of small size neural nodes in human brain that are connected together to construct the considerable networks with complexity relationship. ANN consists of many nodes connecting with lines to compute, learn, and operate specific task. The learning and computation will be done by considering training examples then adjusting weight in each connecting line for the optimum result of predicting value of a target variable. This learning process is self-learning model; that means result can be provided without programming specific rules. The diagram in Fig. 3 shows general architecture of ANN. There are 3 majority layers: input, hidden, and output layers.

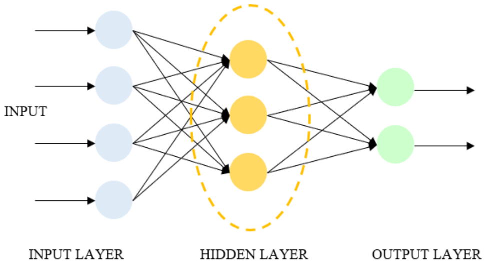


Fig. 3. General structure of simple Artificial Neural Network

1) Input layer consists of input nodes with the number of nodes equals to number of features of a dataset. All of input nodes are connected to hidden layer.

2) Hidden layer consists of hidden nodes with lines connecting to the next level. There can be one or more levels in this hidden layer. Hidden layer is provided information from the nodes in previous hidden layer or input layer.

3) Output layer consists of output nodes. The number of nodes equals to number of values of target variable. The output nodes are always provided the information by the last hidden layer.

H. Performance Evaluation: MAE and RMSE

To evaluate yield prediction performance, we use mean absolute error (MAE) and root mean square error (RMSE) as the measurement tools. MAE and RMSE are typical measurement metrics in yield prediction and many other fields [3]-[5]. The calculation of MAE can be done by averaging differences between actual values of target variable and the predicted values, defined as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{4}$$

where MAE stands for the mean absolute error, n is the numbers of data, y_i is the real value of target variable, and $\hat{y}_i$ is the predicted value made by the model.

RMSE uses the same concept as MAE but the computation is slightly different in that RMSE is to find square root of average of differences between real value of target variable and predicted value power by 2. The formula is provided as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{5}$$

where RMSE stands for the root mean square error.

III. Material and Method

A. Dataset

The dataset used in this research has been collected from the real manufacturing of hard disk drive. The time frame of data collection is 3 months of HDD production. The number of record (or rows) is 1,000,000 rows and number of features (or attributes) is more than 100.

Attributes are the information recorded in the production and test process for every individual HDD unit, as shown in Fig. 4. The 12 attributes given in Fig. 4 are described as follows:

1) Drive serial number (drive SN): This attribute is identification number of each HDD lot. This number is unique for each HDD lot.

2) WEEK: This is the fiscal week that the particular HDD had been assembled.

3) STATUS: This attribute records the status of test process. There are only 2 possible values: Pass and fail. The "pass" status indicates that this HDD passed the test process and be able to be input of the next operation step or ready to ship to customer. The "fail" status means this

HDD is rejected from the test process and must go to either "rework", "retest", "recycle" or "scrap" process according to the debug diagnostic failure symptom.

4) HSA prime-rework status (HSA_PR): This attribute reveals the condition of HSA component. The two possible values of this component are prime and rework. "Prime" means this HSA is the fresh new built component and never been installed in any other HDD before. "Rework" means this HSA is a component that had been installed in another HDD, but that HDD had been rejected in the test process with the HSA labeled as rework. Thus, this HSA is recycled by being rebuilt again in this HDD. (Note that definitions of Prime and Rework are also used in the attributes 5 through 9.)

5) Media prime-rework status (media_PR): This attribute is either the prime or rework condition of media.

6) MBA prime-rework status (MBA_PR): This attribute refers to the prime or rework condition of motor base assembled.

7) VCM prime-rework status (VCM_PR): This attribute describes the prime or rework condition of VCM.

8) TC prime-rework status (TC_PR): This attribute is the prime or rework condition of Top cover.

9) PCBA prime-rework status (PCBA_PR): This attribute is the prime or rework condition of PCBA.

10) DB_Line: This attribute is the identification number of the HDD assembly line.

11) HSA_Line: This attribute is the identification number of the HSA assembly line.

12) PCBA_Line: This attribute reveals the production line for installing PCBA into the HDD.

Drive SN	WEEK	STATUS	HSA_PR	MEDIA_PR	MBA_PR	VCM_PR	TC_PR	PCBA_PR	DB_LINE	PCBA_LINE	HSA_LINE
SN0000001	W01	Pass	Prime	Prime	Prime	Prime	Prime	Prime	DB_1	PCBA_2	HSA_3
SN0000002	W01	Pass	Prime	Prime	Prime	Prime	Prime	Prime	DB_1	PCBA_2	HSA_3
SN0000003	W01	Fail	Prime	Rework	Prime	Prime	Prime	Prime	DB_1	PCBA_2	HSA_3
SN0000004	W01	Fail	Rework	Rework	Prime	Prime	Prime	Prime	DB_1	PCBA_2	HSA_3
SN0000005	W01	Pass	Prime	Prime	Prime	Prime	Prime	Prime	DB_1	PCBA_2	HSA_1
SN0000006	W01	Pass	Prime	Prime	Prime	Prime	Prime	Prime	DB_1	PCBA_2	HSA_1
SN0000007	W01	Pass	Prime	Prime	Prime	Prime	Prime	Prime	DB_1	PCBA_2	HSA_1
SN0000008	W01	Pass	Prime	Prime	Prime	Prime	Prime	Prime	DB_1	PCBA_1	HSA_1
SN0000009	W01	Pass	Prime	Prime	Prime	Rework	Prime	Prime	DB_1	PCBA_2	HSA_1
SN0000010	W01	Pass	Prime	Prime	Prime	Prime	Rework	Rework	DB_2	PCBA_3	HSA_1
SN0000011	W02	Fail	Prime	Rework	Rework	Prime	Rework	Rework	DB_2	PCBA_4	HSA_2
SN0000012	W02	Pass	Prime	Prime	Prime	Prime	Prime	Prime	DB_2	PCBA_5	HSA_2
SN0000013	W02	Pass	Prime	Prime	Prime	Prime	Prime	Prime	DB_2	PCBA_6	HSA_2
SN0000014	W02	Pass	Prime	Prime	Prime	Prime	Prime	Prime	DB_2	PCBA_7	HSA_2
SN0000015	W02	Pass	Prime	Prime	Prime	Prime	Prime	Prime	DB_2	PCBA_8	HSA_2
SN0000016	W03	Pass	Prime	Prime	Prime	Prime	Prime	Prime	DB_2	PCBA_9	HSA_2
SN0000017	W03	Pass	Prime	Prime	Prime	Prime	Prime	Rework	DB_2	PCBA_10	HSA_2
SN0000018	W03	Pass	Prime	Prime	Prime	Prime	Prime	Prime	DB_2	PCBA_11	HSA_2
SN0000019	W03	Pass	Prime	Prime	Rework	Rework	Rework	Prime	DB_2	PCBA_12	HSA_2
SN0000020	W03	Fail	Rework	Prime	Prime	Rework	Prime	Prime	DB_1	PCBA_13	HSA_3

Fig. 4. Example table to show attributes in a hard disk drive manufacturing being grouped by unit.

Beside these main attributes, there are also many other attributes with the diverse meanings and important in a unit of HDD. Total number of attributes used in our experiment is 125.

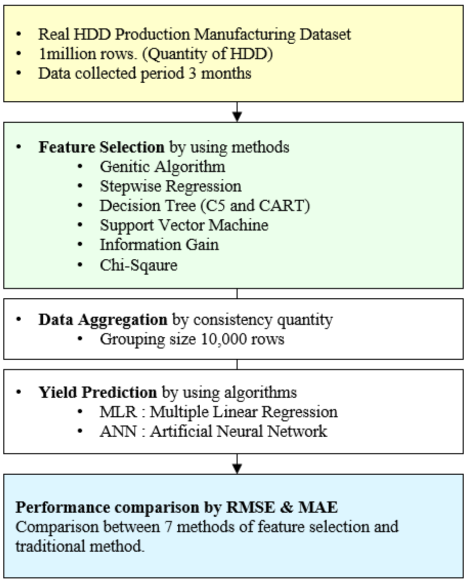
B. Feature Selection Step

Objective of this research is the improvement of yield prediction accuracy by focusing on feature selection part. That means we expect the better performance of yield prediction model built from applying various feature selection techniques.

The feature selection techniques based on machine learning algorithms are C5, CART, SVM, Stepwise Regression, and GA. Moreover, we also include the two techniques based on statistics like correlation filter and chi-square filter. Feature selection by human expert is also used as a baseline for performance comparison. The process engineers select only 5 attributes. These key attributes are considered important based on long-term experience of the engineers.

C. Research Framework

Experimentation steps from data collection, feature selection, data aggregation until yield prediction modeling are schematically displayed in Fig. 5. The 7 feature selection methods are experimented and compared against the key attribute selection method used by the engineers.



- Real HDD Production Manufacturing Dataset
- 1million rows. (Quantity of HDD)
- Data collected period 3 months

- **Feature Selection** by using methods
  - Genitic Algorithm
  - Stepwise Regression
  - Decision Tree (C5 and CART)
  - Support Vector Machine
  - Information Gain
  - Chi-Sqaure

- **Data Aggregation** by consistency quantity
  - Grouping size 10,000 rows

- **Yield Prediction** by using algorithms
  - MLR : Multiple Linear Regression
  - ANN : Artificial Neural Network

Performance comparison by RMSE & MAE
Comparison between 7 methods of feature selection and traditional method.

Fig. 5. Research framework

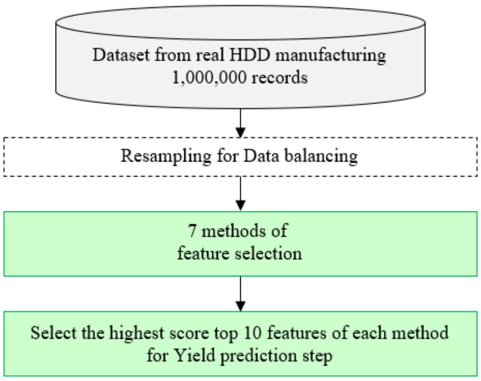

Dataset from real HDD manufacturing
1,000,000 records
Resampling for Data balancing
7 methods of
feature selection
Select the highest score top 10 features of each method
for Yield prediction step

Fig. 6. Research workflow for the part of feature selection

D. Research Workflow

Flow chart diagram of Fig. 6 depicts the research workflow in a specific part of feature selection. The experiment starts by resampling to make data balancing. Data balancing is necessary because from observing the original data containing totally 1,000,000 records, we found that the target class (status) of dataset is skewed with higher number of pass than fail. The imbalance ratio between majority (pass) and minority (fail) is about 28:1. After data balancing step, each of the 7 feature selection methods are applied to select features from 125 attributes. These methods return the result by ranking the important factors in descending order. Then, we select the top 10 attributes of each method to be used further in the next step of modeling to create a yield prediction model.

In yield prediction modeling, the process starts by aggregating the 1,000,000 data records to become 100 rows in which each row contains 10,000 records. This aggregation steps is for accuracy improvement as we obse5rved from our preliminary experiments. After data aggregation step, we obtain new 7 datasets, each of which is a dataset with 100 rows and 10 attributes that are selected from 7 methods of feature selection. That means these new datasets have 100 rows aggregated from the original 10,000 record with ten attributes that can be different from one dataset to the others because they are selected with different methods.

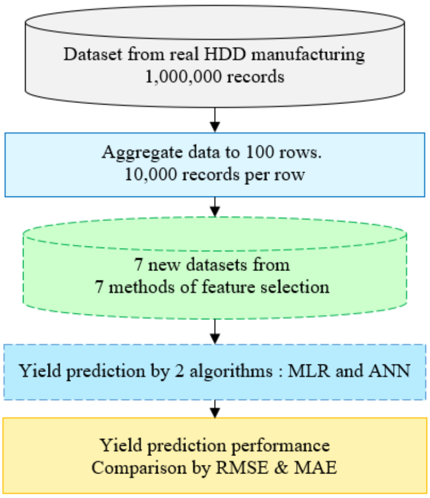

Dataset from real HDD manufacturing
1,000,000 records
Aggregate data to 100 rows.
10,000 records per row
7 new datasets from
7 methods of feature selection
Yield prediction by 2 algorithms : MLR and ANN
Yield prediction performance
Comparison by RMSE & MAE

Fig. 7. Research workflow for the part of data aggregation and yield prediction modeling

After that, the learning algorithms MLR and ANN are applied to create yield prediction models. Finally, yield prediction performances are compared among 7 methods of feature selection and human selection method done by the process engineers. All of these steps are depicted as a workflow diagram and shown in Fig. 7.

IV. Experimental Results

The main focus of our experimentation is to study performances of 7 different methods for feature selection that should adopted for data preparation step prior to building the model to predict pass/fail of the assembled HDD units. The accuracy of pass/fail prediction is important for yield computation. The more accurate Pass/Fail prediction, the more precise yield estimation.

The 7 methods for feature selection used in this work are C5, CART, SVM, GA, stepwise regression, chi-square, and Information Gain. The first five algorithms are also learning algorithms, whereas the last two are only for feature ranking and selection. Thus, we firstly, apply the 7 algorithms for selecting the top-10 features anticipating to contribute the most toward yield prediction through the accurate forecasting of the HDD status as either pass or fail. The five algorithms that are both capable of feature selection and learning to build model are applied for both jobs. Their accuracies are computation time are reported and shown in Table II. The two algorithms (chi-square and information gain) that can only be used for feature selection are reported just for their computation time.

It can be observed from the results that the three models with highest pass/fail prediction accuracy are those built by C5, CART and SVM, respectively. In terms of computation time, chi-square and information gain show outstanding shorter time (1 second). GA takes the longest time of feature selection step at 4,620 seconds.

Table II: Accuracy and Computation Time of 7 Methods in Feature Selection Step

Feature Selection Methods	Accuracy (%)	Time (sec)
C5	72.38	646.8
CART	66.57	31.7
SVM	64.62	1,216.8
Stepwise Regression	64.76	56.5
Genetic Algorithm	61.94	4,620.2
Chi-Square	-	1.0
Information Gain	-	1.0

For the next part of our experimentation based on the dataset with selected features, yield has been computed as: (quantity of pass units) / (quantity of input units). Actual yield values and predicted yields made by the learning algorithms (MLR and ANN) are compared and the prediction errors are shown in Table III. At this step, yield prediction results based on the five key features selected by human experts are also shown in the first row of the table as a baseline for performance comparison.

TABLE III: ACCURACY AND COMPUTATION TIME OF 7 METHODS IN FEATURE SELECTION STEP

Feature selection method	Yield prediction algorithm	Test data	
		RMSE	MAE
Human Engineers	Traditional	0.01700	0.01400
C5	MLR	**0.00866**	**0.00605**
	ANN	0.01707	**0.01263**
CART	MLR	0.24105	0.05913
	ANN	**0.01630**	**0.01251**
SVM	MLR	0.02037	**0.01247**
	ANN	0.01864	**0.01384**
Stepwise Regression	MLR	0.10326	0.02842
	ANN	0.01851	**0.01306**
Genetic Algorithm	MLR	**0.00732**	**0.00559**
	ANN	0.01706	**0.01269**
Chi-Square	MLR	**0.00821**	**0.00690**
	ANN	0.01707	**0.01262**
Information Gain	MLR	**0.00821**	**0.00690**
	ANN	0.01707	**0.01262**

TABLE IV: ERROR REDUCTION FROM TRADITIONAL ENGINEERING METHOD

Feature selection and model building scheme	Error reduction	
	RMSE	MAE
C5 with MLR	-49%	-57%
C5 with ANN	0%	-10%
CART with MLR	1318%	322%
CART with ANN	-4%	-11%
SVM with MLR	20%	-11%
SVM with ANN	10%	-1%
Stepwise with MLR	507%	103%
Stepwise with ANN	9%	-7%
GA with MLR *	-57%	-60%
GA with ANN	0%	-9%
Chi-Square with MLR	-52%	-51%
Chi-Square with ANN	0%	-10%
Information Gain with MLR	-52%	-51%
Information Gain with ANN	0%	-10%

In terms of RMSE, statistical feature selection methods like chi-square and information gain when modeling with MLR perform better than feature selection made by human engineers. However, when building the model with ANN, both methods are as good as the human expert. C5 and GA are also comparable to traditional method when using ANN yield prediction model.

It can be seen from the results that feature selection with GA and then building the model with MLR yield the best result with least RMSE value at 0.00732. Comparing to traditional method and feature selection made by human expert, the combination of GA and MLR can significantly improve yield prediction performance with error reduction around 57% (as shown in Table IV).

When considering from the MAE metric with error reduction computed by using traditional method with human selected features as summarized in Table IV, it can be seen that almost all machine learning based and statistical based modeling methods with the base value of MAE = 0.014. This is except the two combinations, CART + MLR and stepwise regression + MLR that perform worse than human feature selection + traditional method. The best yield prediction scheme in terms of MAE metric is Genetic Algorithm with MLR prediction model.

V. CONCLUSION

This paper introduces the novel idea of applying machine learning and statistical analysis techniques in feature selection part to improve performance of yield prediction in the Hard Disk Drive (HDD) manufacturing. The assumption of this research is that the number of features from HDD manufacturing is typically numerous and thus prediction performance can be lessen by the shadow of too many features. We propose that proper feature selection technique can help improving yield prediction by selecting only key important features. Efficiency of this proposal has been confirm through experiments with the real-world data collected from HDD manufacturing containing 1 million records and 125 attributes. The experiments are done by applying 7 methods of feature selection and the yield prediction models are built from the two learning algorithms.

The experimental results demonstrate that in terms of RMSE metric, the 4 from 7 feature selection methods in combination with the MLR learning algorithm can help improving yield prediction performance. In terms of MAE metric, all 7 feature selection methods in combination with the ANN learning algorithm can improve yield prediction. The best combination is GA and MLR can improve performance when compared against traditional method that required human engineers to select key features the improvement is as high as 57%. However, the trade-off from using GA is the long computation time. These results lead to conclusion that the proposed novel idea of combining feature selection technique with powerful learning algorithm can help improving yield prediction performance in the real application of HDD manufacturing.

RECOMMENDATION

The dataset used in this research had been collected from 3 months of production timeframe in the steady and maturity performance phase. Yield computation of this dataset gives the results that are quite stable with low fluctuation. In the future, researchers and engineering experts in HDD manufacturing agree to make some challenging advancement by using dataset of “developing phase” instead of “maturity phase”. This challenge can gain more benefit because of the successful result in data of developing phase can help manager to prepare good action in mass production of maturity phase.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

The first author is responsible for designing the research framework, organizing the experimentation steps and preparing the manuscript. The second author advice and help in developing programs with R and Python languages. The third author helps editing the manuscript and validating the research steps. The last author helps confirming the experimental results and discussing the future research trends.

References

[1] R. Wood, "Future hard disk drive systems," *Journal of Magnetism and Magnetic Materials*, vol. 321, no. 6, pp. 555-561, 2009.

[2] V. Kasavajhala "Solid state drive vs. hard disk drive price and performance study," *A Dell Technical White Paper*, Dell PowerVault Storage Systems, 2011, pp. 1-13.

[3] H. Lee, C. O. Kim, H. H. Ko, and M. Kim, "Yield prediction through the event sequence analysis of the die attach process," *IEEE Trans. on Semiconductor Manufacturing*, vol. 28, no. 4, pp. 563-570, 2015.

[4] J. Li, X. Ji, Y. Jia, *et al*., "Hard drive failure prediction using classification and regression trees," in *Proc. 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, 2014, pp. 383-394.

[5] T. Yuan, S. Z. Ramadan, and S. J. Bae, "Yield prediction for integrated circuits manufacturing through hierarchical Bayesian modeling of spatial defects," *IEEE Trans. on Reliability*, vol. 60, no. 4, pp. 729-741, 2011.

[6] K. N. Malitski, "Method for shipment planning/scheduling," Patent, US8244645, 2012.

[7] A. K. R. Katta and R. Allgor, "Heuristic methods for customer order fulfillment planning," Patent, US8352382, 2013.

[8] M. Braglia, D. Castellano, M. Frosolini, and M. Gallo, "Overall material usage effectiveness (OME): A structured indicator to measure the effective material usage within manufacturing processes," *Production Planning & Control*, vol. 29, no. 2, pp. 143-157, 2018.

[9] J. Shi, G. Zhang, and J. Sha, "Optimal production planning for a multi-product closed loop system with uncertain demand and return," *Computers & Operations Research*, vol. 38, no. 3, pp. 641-650, 2011.

[10] A. P. Rastogi, J. W. Fowler, W. M. Carlyle, O. M. Araz, A. Maltz, and B. B üke, "Supply network capacity planning for semiconductor manufacturing with uncertain demand and correlation in demand considerations," *International Journal of Production Economics*, vol. 134, no. 2, pp. 322-332, 2011.

[11] S. M. Liozu and A. Hinterhuber, "Industrial product pricing: a value-based approach," *Journal of Business Strategy*, vol. 33, no. 4, pp. 28-39, 2012.

[12] D. A. Patterson and J. L. Hennessy, *Computer Organization and Design ARM Edition: The Hardware Software Interface*, Morgan Kaufmann, 2016.

[13] J. S. Domingo. (January 25, 2019). SSD vs. HDD: What's the difference. *PC Magazine*. [Online]. Available: https://sea.pcmag.com/storage/1526/ssd-vs-hdd-whats-the-difference

[14] N. U. Mustafa, A. Armejach, O. Ozturk, A. Cristal, and O. S. Unsal, "Implications of non-volatile memory as primary storage for database management systems," in *Proc. Int. Conf. on Embedded Computer Systems: Architectures, Modeling and Simulation (SAMOS)*, 2016, pp. 164-171.

[15] G. G. Foisy, N. E. Larson, S. Narayan, A. F. Sancheti, and J. Edwards, "Head stack assembly for a disk drive having a unitary molded plastic E-block, " Patent, US6061206, 2000.

[16] K. Takaishi, Y. Uematsu, T. Yamada, M. Kamimura, M. Fukushi, and Y. Kuroba, "Hard disk drive servo technology for media-level servo track writing," *IEEE Trans. on Magnetics*, vol. 39, no. 2, pp. 851-856, 2003.

[17] A. A. Mamun, G. Guo, and C. Bi, *Hard Disk Drive: Mechatronics and Control*, CRC press, 2017.

[18] T. R. Simon, L. Cong, Y. Zhai, Y. Zhu, and F. Zhao, "A semi-automatic system for efficient recovery of rare Earth permanent magnets from hard disk drives," *Procedia CIRP*, vol. 69, pp. 916-920, 2018.

[19] V. W. Santini and A. D. Little, "Disk drive including a printed circuit board assembly and a PCBA shield with tabs engaged in slots of a disk drive base," Patent, US7271978, 2007.

[20] H. Turabieh, M. Mafarja, and X. Li, "Iterated feature selection algorithms with layered recurrent neural network for software fault prediction," *Expert Systems with Applications*, vol. 122, pp. 27-42, 2019.

[21] A. Suppers, A. J. V. Gool, and H. J. Wessels, "Integrated chemometrics and statistics to drive successful proteomics biomarker discovery," *Proteomes*, vol. 6, no. 2, pp. 20, 2018.

[22] H. Liu and M. Zhou, "Decision tree rule-based feature selection for large-scale imbalanced data," in *Proc. 26th Wireless and Optical Communication Conference*, 2017, pp. 1-6.

[23] H. Frohlich, O. Chapelle, and B. Scholkopf, "Feature selection for support vector machines by means of genetic algorithm," in *Proc. 15th IEEE International Conference on Tools with Artificial Intelligence*, 2003, pp. 142-148.

[24] J. Huang, Y. Cai, and X. Xu, "A hybrid genetic algorithm for feature selection wrapper based on mutual information," *Pattern Recognition Letters*, vol. 28, no.13, pp. 1825-1844, 2007.

[25] M. Anbarasi, E. Anupriya, and N. C. S. N. Iyengar, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm," *International Journal of Engineering Science and Technology*, vol. 2, no. 10, pp. 5370-5376, 2010.

[26] G. Stein, B. Chen, A. S. Wu, and K. A. Hua, "Decision tree classifier for network intrusion detection with GA-based feature selection," in *Proc. 43rd Annual Southeast Regional Conference*, vol. 2, 2005, pp. 136-141.

[27] R. Pandya and J. Pandya, "C5.0 algorithm to improved decision tree with feature selection and reduced error pruning," *International Journal of Computer Applications*, vol. 117, no. 16, pp. 18-21, 2015.

[28] G. A. Seber and A. J. Lee, *Linear Regression Analysis*, John Wiley & Sons, 2012.

[29] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, John Wiley & Sons, 2012.

[30] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied Linear Statistical Models*, Chicago, 1996.

[31] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," *International Journal of Forecasting*, vol. 14, no. 1, pp. 35-62, 1998.

[32] Y. T. Chae, R. Horesh, Y. Hwang, and Y. M. Lee, "Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings," *Energy and Buildings*, vol. 111, pp. 184-194, 2016.

[33] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Proc. Advances in Neural Information Processing Systems*, 2016, pp. 523-531.

A. Hirunyawanakul is a Ph.D. student, School of Computer Engineering, Suranaree University of Technology (SUT), Thailand. She received her B.E. and M.E. in computer engineering from Suranaree University of Technology, Thailand, in 2006 and 2014. Her research of interest includes Data Mining, Machine Learning, and Artificial Intelligence.

N. Kaoungku is currently a lecturer at School of Computer Engineering, Suranaree University of Technology (SUT), Thailand. He received his bachelor, master, and doctoral degrees in Computer Engineering from SUT in 2012, 2013, and 2015, respectively. His current research work includes Data Mining, Knowledge Engineering, and Semantic Web.

N. Kerdprasop is an associate professor and the head of Data Engineering Research Unit, School of Computer Engineering, Suranaree University of Technology (SUT), Thailand. She received her B.S. in radiation techniques from Mahidol University, Thailand, in 1985, M.S. in computer science from the Prince of Songkla University, Thailand, in 1991 and Ph.D. in computer science from Nova Southeastern University, U.S.A., in 1999. Her research of interest includes Data Mining, Artificial Intelligence, Logic and Constraint Programming.

K. Kerdprasop is an associate professor at the School of Computer Engineering, Chair of the School, and the head of Knowledge Engineering Research Unit, SUT. He received his bachelor degree in Mathematics from Srinakarinwirot University, Thailand, in 1986, MS in computer science from the Prince of Songkla University, Thailand, in 1991 and Ph.D. in computer science from Nova Southeastern University, U.S.A., in 1999. His current research includes Machine Learning and Artificial Intelligences.