IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 February 2024, accepted 20 February 2024, date of publication 23 February 2024, date of current version 29 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3369443

RESEARCH ARTICLE

Building a Rule-Based Expert System to Enhance the Hard Disk Drive Manufacturing Processes

SUPPAKRIT KIRDPONPATTARA[1], PITIKHATE SOORAKSA[2], AND VEERA BOONJING[3]

[1]School of International and Interdisciplinary Engineering Programs, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Ladkrabang, Bangkok 10520, Thailand
[2]Department of Robotics and AI, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Ladkrabang, Bangkok 10520, Thailand
[3]Department of Computer Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Ladkrabang, Bangkok 10520, Thailand

Corresponding author: Suppakrit Kirdponpattara (63601230@kmitl.ac.th)

ABSTRACT The manufacturing of hard disk drives involves the intricate assembly of numerous components, making the testing process time-consuming and resource intensive. To optimize the manufacturing process and increase testing efficiency, the development of a rule-based expert system is proposed. This system leverages predictive models constructed from assembly process data to identify potentially defective hard drives before undergoing extensive testing. By preemptively identifying defects, this approach substantially reduces testing time and enhances tester capacity. Given the categorical and imbalanced nature of assembly data, Decision Trees are employed as the prediction model. Specifically, three Decision Tree algorithms are explored: ID3, C4.5, and CART. In addition, four feature selection techniques, namely Information Gain, Gain Ratio, Chi-Square, and Symmetrical Uncertainty, are utilized to identify high-impact features. Our experimental findings reveal that Information Gain coupled with the C4.5 algorithm yields the most favorable results in terms of prediction accuracy, modeling efficiency, and rule generation. Moreover, our study establishes that setting the failure probability threshold between 0.15 and 0.70 provides the shortest total test time for the proposed process, as supported by a 95% confidence level. This achievement represents a statistically significant enhancement compared with the existing manufacturing process.

INDEX TERMS Decision tree, defect prediction, expert system, feature selection, hard disk drive manufacturing.

NOMENCLATURE

χ^2	Chi-Square.
ANN	Artificial Neural Network.
BPNN	Back Propagation Neural Network.
C4.5	Improved version of ID3.
CART	Classification and Regression Tree.
CT	Classification Tree.
DT	Decision Tree.
eFMEA	Extended Failure Mode and Effects Analysis.
ECLPS	Enhanced Common Lisp Production System.
FN	False Negative.
FP	False Positive.
FS	Feature Selection.
GA	Genetic Algorithm.
GBT	Gradient Boosting Tree.
GR	Gain Ratio.
HABs	Harmful Algal Blooms.
HDD	Hard Disk Drive.
HGA	Head Gimbal Assembly.
HSA	Head Stack Assembly.
ID3	Iterative Dichotomiser 3.
IG	Information Gain.
MBA	Motor Base Assembly.
MCFs	Microwave Cavity Filters.
MLR	Multiple Linear Regression.

The associate editor coordinating the review of this manuscript and approving it for publication was Amin Zehtabian.

NB	Naive Bayes.
NICU	Neonatal Intensive Care Unit.
NN	Neural Network.
NRO	Nutrition Recommendation Ontology.
OPC-UA	Open Platform Communications Unified Architecture.
ORF	Online Random Forest.
PCBA	Printed Circuit Board Assembly.
RAT	Rank-sum test Attribute.
RF	Random Forest.
RNN	Recurrent Neural Network.
SMART	Self-Monitoring, Analysis, and Reporting Technology.
SU	Symmetrical Uncertainty.
SVM	Support Vector Machine.
TN	True Negative.
TP	True Positive.
VCM	Voice Coil Motor.

I. INTRODUCTION

The production process of Hard Disk Drives (HDDs) involves assembling the device from various components using different assembling machines. Once assembled, each HDD undergoes a testing operation, which is critical to identifying defects and ensuring that only fully functional HDDs are shipped to customers. The detail process is as shown in Figure 1. Testing each hard drive unit is a costly and time-consuming process because each unit is composed of numerous components, and testing must ensure that each of these components functions properly and operates in sync with the other components [1], [2], [3], [4]. Therefore, testing a product with a high storage capacity, e.g., 20 TB HDD, takes at least one month, regardless of whether it is defective or not. In fact, a known defective HDD can be removed from the normal full testing process. This could improve the manufacturing process by increasing the test capacity. A rule-based expert system with the objective of optimizing the manufacturing process of hard disk drives is proposed in this study. As illustrated in Figure 2, the system uses predictive models to identify potentially defective drives before they undergo extensive testing. This is achieved by classifying hard drives into passers or defectives based on assembly data. A hard drive predicted as defective is subjected to a tailored, shorter testing process. This process, which involves selecting the most relevant test operations from a complete list, quickly identifies defects. This approach not only expedites defect detection but also reduces testing time and resource consumption, thereby enhancing the manufacturing process. The effectiveness of this approach hinges on the accuracy of the prediction model. The model, which relies solely on information about the components and assembly machines used, applies feature selection techniques and a decision tree algorithm to create minimal and interpretable classification rules. These components and machines may be supplied by different vendors. Figure 3 further elaborates on the workings of the expert system. It takes the assembly data as input and

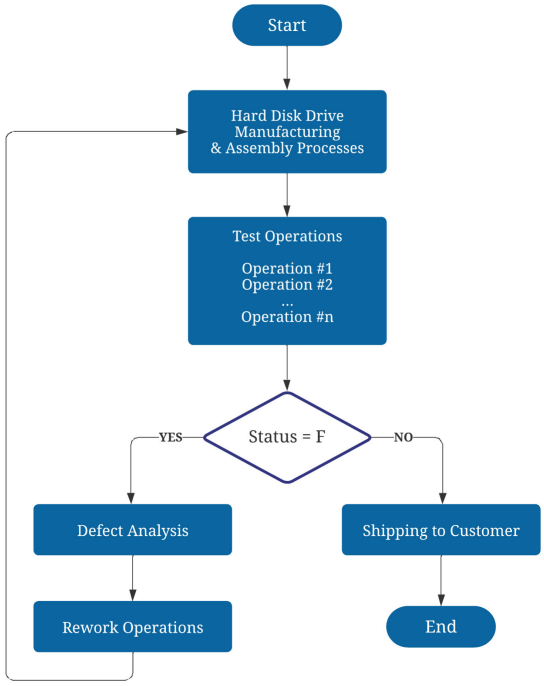

FIGURE 1. A flow chart of the current production process of hard disk drives, showing the assembly and testing operations.

generates rules for defect prediction. These rules are then used to assign a failure probability to each hard drive. If the failure probability exceeds a predefined threshold, the hard drive is classified as defective; otherwise, it is classified as a passer.

This study is confronted with two primary challenges in constructing a prediction model. They are (1) all input features to the model are categorical, and (2) there are many input features. A Decision Tree (DT) is a well-understandable binary classifier, and it is the most popular used for categorical data [5]. It often performs well on imbalanced data because its hierarchical structure allows it to learn signals from both classes. Therefore, this study chooses the DT as the prediction model. This study evaluates the ID3, C4.5, and CART algorithms [6] to identify the most accurate one. The large number of input features may cause DTs learning from all input features to suffer from low predictive accuracy, due to unrelated input features to the target class. To address this problem, feature selection (FS) [7] is proposed by exploring all combinations of four FS techniques (Information Gain, Gain Ratio, Chi-Square, and Symmetrical Uncertainty) with the three DTs mentioned above. The best combination is determined in terms of prediction accuracy, modeling time, and the number of rules. The study further investigates this best combination to find its optimal parameters that minimize the total test time of the proposed process.

In summary, this study embarks on the development of a rule-based expert system aimed at enhancing the manufacturing processes of Hard Disk Drives. Through the utilization of predictive modeling, tailored testing, and

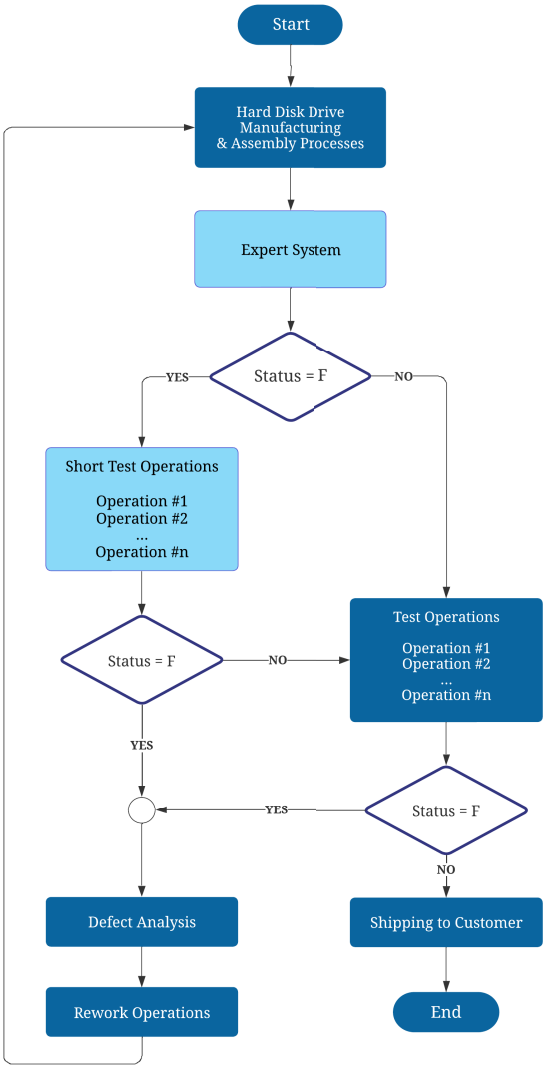

FIGURE 2. A flow chart of the proposed production process of hard disk drives, showing the expert system for defect prediction and the tailored testing process.

feature selection, the aim is to streamline defect detection, reduce resource consumption, and ultimately optimize HDD production.

The rest of this paper is organized as follows. Section II provides information about the HDD product and its assembly process, as well as the background and related works on defect prediction, decision tree algorithms, and expert systems. Furthermore, it explains the methodology of the rule-based expert system, which uses feature selection techniques and decision tree algorithms to classify HDDs into passers or defectives based on assembly data. Section III presents the results of the experiments, which evaluate the performance of the expert system in terms of accuracy, modeling time, rule generation, and test time reduction. Section IV discusses the ethical aspects of the research. Section V concludes the article and suggests some directions for future work.

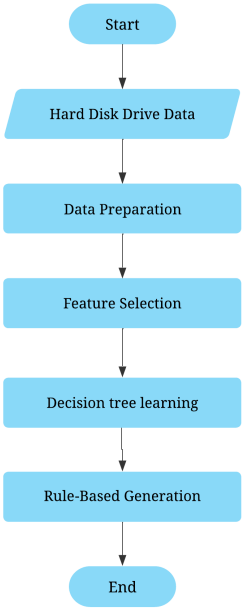

FIGURE 3. A flow chart of the expert system for defect prediction, showing the input, output, and components of the system.

II. MATERIALS AND METHODS

This section provides the background and related works pertaining to the present study. The initial part outlines the HDD product and its assembly process. The following parts present illustrative examples on FS techniques and DT algorithms employed in this study. Finally, the last part of this section provides an overview of the relevant works in this area.

A. HARD DISK DRIVE

The HDD is a non-volatile digital data storage device that utilizes magnetic storage technology to store and retrieve data on durable material disks [8]. The production of HDD involves the assembly of hardware components sourced from various vendors, performed by each production machine. As illustrated in Figure 4, the HDD comprises six major components [9]. These components include:

Base Deck/Motor base assembly (MBA) refers to the aluminum casing that safeguards all internal hardware components of the HDD from external factors [10].

Spindle Motor is responsible for the control and management of the rotational speed of the disks [11].

Disk Platters are the components that store the data, comprising platinum group metals deposited on a substrate [12].

Head stack assembly (HSA) is composed of the Actuator Arm and Head Gimbal Assembly (HGA), which incorporates the assembly of the slider and the suspension.

The HSA is responsible for synchronizing the movement of the heads to the appropriate reading/writing position, in conjunction with the motor rotation speed [13].

Voice Coil Motor (VCM) is an electromechanical linear motor responsible for positioning the read/write heads by moving HSA [14].

FIGURE 4. A diagram of the hard disk drive components, showing the base deck, spindle motor, disk platters, head stack assembly, voice coil motor, and printed circuit board assembly.

Printed Circuit Board Assembly (PCBA) is a controller board that controls HDD and provides an interface between HDD and the computer [15].

All components will be assembled with the order of process flow for each production machine. Hence, the manufacturing of each HDD component sourced from various vendors and produced by various machines will determine the segregation of HDD into defect or passer.

B. DECISION TREE

A decision tree is a graphical representation of either a classification or regression problem. Each node in this tree corresponds to a feature, each branch represents a decision rule, and each leaf node represents an outcome. The decision tree has the capability to recursively partition data into smaller subsets based on feature values until a stopping criterion is met. One of the advantages of using a decision tree is its ease of interpretation and explanation. It also has the ability to handle both categorical and numerical data. However, decision trees can be prone to overfitting, sensitive to noise and outliers, and unstable to small changes in the data. This section provides a description of three popular decision tree algorithms, namely ID3, C4.5, and CART.

1) ITERATIVE DICHOTOMISER 3 (ID3)

This algorithm, published in 1986 [16], is a well-known decision tree technique. It uses Information Gain (IG) to choose the best features for modeling. IG measures the reduction in Entropy, which is the uncertainty or randomness of the data, after splitting the data by a feature.

The formula for IG is:

$$IG(T,f) = Entropy(T) - \sum_{v \in Values(f)} P(T|f=v) \times Entropy(T|f=v) \quad (1)$$

Entropy, which is calculated using the following formula, measures the level of disorder or randomness in the dataset:

$$Entropy(T) = -\sum_{c \in Classes(T)} P(c) \times log_2 P(c) \quad (2)$$

where T is the dataset, f is the feature, $P(T|f = v)$ is the probability of each value in feature f, and $P(c)$ is the probability of each class in dataset T.

ID3 can handle categorical data well, but it has some limitations. It does not work with missing values and it may overfit the data if the tree grows too large.

2) C4.5

The C4.5 decision tree algorithm, an extension of ID3, was introduced in 1993 [17]. It is capable of handling both categorical and numerical data. C4.5 uses the gain ratio (GR) for feature selection, which mitigates the bias of information gain towards features with many values. The GR is defined as the ratio of information gain and split information, as shown in the following formulas:

$$GR(T,f) = \frac{IG(T,f)}{SplitInfo(T,f)} \quad (3)$$

$$SplitInfo(T,f) = -\sum_{v \in Values(f)} P(T|f=v) \times \log_2 P(T|f=v) \quad (4)$$

where T is the dataset, f is the feature, $P(T|f = v)$ is the probability of each value in feature f, and $IG(T,f)$ is the information gain of the feature.

The algorithm also supports pruning the tree to avoid overfitting and can handle missing values by assigning them to the most common class or value.

3) CLASSIFICATION AND REGRESSION TREE (CART)

The CART, introduced in 1984 [18], is a decision tree algorithm that uses the Gini Index as a criterion for feature selection. The Gini Index measures the impurity of a node, which is the probability of a randomly chosen sample being

incorrectly classified. The Gini Index is calculated using the following formulas:

$$GiniIndex(T|f) = \sum_{v \in Values(f)} \frac{|T|f = v|}{|T|} \times Gini(T|f = v) \tag{5}$$

$$Gini(T|f = v) = 1 - \sum_{c \in Classes(T|f=v)} P(c)^2 \tag{6}$$

where T is the dataset, f is the feature, $|T|$ is the total count of items in the dataset, $|T|f = v|$ is the count of items in the dataset T where the feature f equals the value v, $Gini(T|f = v)$ is the Gini of each value in feature f, and $P(c)$ is the probability of each class in dataset T.

CART aims to find the feature and the split point that minimize the weighted average of the Gini Index of the child nodes. CART can handle both categorical and numerical data, and it only allows binary splits. It can also deal with missing values by assigning them to the most similar subset or using surrogate splits, which are alternative splits that mimic the best split as closely as possible.

C. EXPERT SYSTEM

An expert system is a software application that simulates the problem-solving skills of a human specialist in a specific domain. It consists of a knowledge base, which stores facts and rules about the domain, and an inference engine, which applies logical reasoning to infer solutions or suggestions [19]. One of the benefits of using an expert system is its ability to provide consistent and reliable answers to complex problems, handling uncertainty and incomplete information through techniques such as probability.

In 2019, Viktor proposed a method for assessing the state of complex systems under random non-systematic influences, based on subjective probabilities derived from expert opinion. The author developed an algorithm for forming an expert team and a method for constructing expert systems using Bayes' theorem. They presented a software implementation of the expert system in C++ and an application to fault diagnosis of telecommunication networks, demonstrating its capability to estimate the subjective probability of the state of a complex system and support decision-making [20].

For fuzzy logic, in 2022, Leyu et al. proposed a dynamic-attention-based heuristic fuzzy expert system for tuning microwave cavity filters (MCFs). The method includes multiple evaluation functions, a dynamic-attention-based expert system, and a heuristic fuzzy logic system, achieving automatic tuning with high applicability, accuracy, and efficiency for MCFs. Simulations and experiments validated the method's effectiveness, applicability, and practicality [21].

Foni et al. proposed an automated algae species identification system in 2022, utilizing ontology and certainty factors to assist and validate expert judgment. Tested on 60 samples of 20 common harmful algal bloom (HAB) species in Lampung Bay and Jakarta Bay, Indonesia, the system achieved an accuracy of 73.33% and high agreement with expert identification on six algae species. This system could be used as an alternative tool for rapid algal identification or as part of an early warning system for HABs [22].

However, expert systems have drawbacks. They can be hard to maintain and update, as the knowledge base may become outdated or inconsistent over time. Additionally, they may lack common sense, creativity, or flexibility, being limited by the rules and facts in the knowledge base.

1) RULE-BASED EXPERT SYSTEMS

Rule-based expert systems represent a significant branch of expert systems, using IF-THEN rules to encapsulate expert knowledge in a specific domain. Renowned for their simplicity and interpretability, rule-based expert systems find applications in various fields, including medicine. For example, Ravneet et al. presented an ontology and rule-based clinical decision support system for personalized nutrition recommendations in the neonatal intensive care unit (NICU). The authors developed the Nutrition Recommendation Ontology (NRO), achieving a 98% accuracy rate in validation [23].

In industrial settings, a rule-based expert system for fault detection and diagnosis leverages an extended Failure Mode and Effects Analysis (eFMEA). A methodology for digitalizing the eFMEA and generating rules for the expert system is proposed, allowing communication with equipment using OPC-UA protocol and handling multi-fault scenarios. A web application visualizes fault detection and diagnosis results, providing a clear description of the approach, implementation, benefits, limitations, and future work directions [24].

For manufacturing process control, [25] presented a rule-based simulator for a semiconductor manufacturing line, written in Enhanced Common Lisp Production System (ECLPS), a knowledge-based language. The simulator uses a single-rule template to move product lots through various process steps, customized with data for each step, route, lot, tool, etc. The model is simple, flexible, and maintainable, running daily at the IBM semiconductor manufacturing plant in Yasu, Japan, on three different semiconductor manufacturing lines.

As mentioned, expert systems and rule-based expert systems have been used in various fields such as medicine, engineering, industrial, education, and manufacturing. In the context of hard disk drive manufacturing, expert systems can improve the production process, reduce defects, and enhance quality. For instance, an expert system can use assembly data to identify defective drives before extensive testing, as proposed in this paper.

D. RELATED WORKS

There are two categories of studies for predicting faults issues in the HDD industry: hard drive failure and manufacturing defects. The first employs operating data of hard drives named SMART (Self-Monitoring, Analysis, and Reporting Technology) to forecast the status of HDD. Well-known data sources for this kind of investigation include Backblaze and

TABLE 1. Studies on hard drive failure.

Previous Related Research	Data Source	Machine Learning Algorithm	Feature Selection Technique	Number of SMART Attributes	Failure Detection Rate
Li (2014) [26]	Baidu W and Q	DT (CART) and BPNN	RAT, rank-sum test, and z-scores	12-19 features	95%
Xu (2016) [27]	Baidu W, S, and M	RNN, CT, and Binary NN	RAT, rank-sum test, and z-scores	10 features	97%
Aussel (2017) [28]	Backblaze (2014)	SVM, RF and GBT	RAT, rank-sum test, z-scores, and quantile function	9 features	95%
Xiao (2018) [29]	Backblaze (2013-2014)	ORF	Rank-sum test	19 features	98% (Dataset: STA) 85% (Dataset: STB)
Wasim (2020) [30]	Backblaze (Q1-Q3 2015)	NB and SVM	Genetic Algorithm	42 features	98.4% (ML Algorithm: NB) 40% (ML Algorithm: SVM)

Baidu. Table 1 summarized studies on hard drive failure [26], [27], [28], [29], [30].

The second employs manufacturing data to predict defects. For the last decade, only a limited number of machine learning algorithms have been applied to assist HDD manufacturing. These applications focus on improving (1) the yield in hard drive manufacturing process [31] and (2) the performance of HDD yield prediction [32]. The objective of the study such as [31] is to improve the yield in hard drive manufacturing process. To achieve this, a decision tree technique to identify controllable parameters that could reduce for HDD defects. The parameters were classified into three categories: uncontrollable, controllable, and dependent. The decision tree algorithm (C4.5) was applied to determine the model with the highest accuracy. The model rules were used to select the controllable parameters, which were then adjusted to increase the number of passers. The results of the experiment showed an increase in passers. However, the experiments were only conducted in a test environment, making it difficult to determine how well the method would perform in real-world scenarios. In a study focused on improving the performance of HDD yield prediction in the manufacturing process [32], researchers utilized machine learning algorithms and feature selection techniques. They applied seven FS techniques, including Decision Tree (C5 and CART), Support Vector Machine (SVM), Stepwise Regression, Genetic Algorithm (GA), Chi-Square, and Information Gain, to select top 10 features for modeling. Multiple Linear Regression (MLR) and Artificial Neural Networks (ANN) algorithms were used for modeling. The study concluded that the best prediction performance was achieved with the combination between GA and MLR, but GA had the longest computation time.

In this research, a practical issue encountered by manufacturers of HDDs is tackled, specifically the reduction of defective drive test time and the augmentation of test equipment capacity. A rule-based expert system that utilizes data from the assembly process to pinpoint defects is proposed, with the caveat that the accuracy of these predictions is a determining factor.

Several advantages over previous research in this field are offered by this methodology. Firstly, testing time and resource consumption can be decreased by identifying defective hard drives prior to comprehensive testing. Secondly, HDD production can be enhanced by expanding the test capacity and refining the defect detection process. Lastly, minimal and interpretable classification rules can be formulated by employing feature selection techniques and decision tree algorithms.

However, some limitations are also presented by this approach that warrant further investigation. Firstly, the model is predicated on a single production timeframe and may not extrapolate well to other timeframes or HDD products. Secondly, the effectiveness of the model is contingent on the quality and availability of the assembly data, which can fluctuate depending on the vendors and machines involved in the production process. Thirdly, all factors influencing the defect probability, such as environmental conditions, human errors, or hardware failures, might not be accounted for by the model.

When compared to previous studies that concentrated solely on hard drive failure prediction or HDD yield prediction improvement, the model is found to be more comprehensive and effective as it encompasses both aspects of the manufacturing process. In contrast to previous studies that employed complex or computationally intensive machine learning algorithms or feature selection techniques, the model is more practical and robust, utilizing simple and efficient methods capable of handling categorical and imbalanced data. Finally, unlike previous studies that used a fixed number of features or a predefined threshold for defect prediction, the model is more flexible and scalable, permitting feature selection and threshold optimization based on the data and performance criteria.

E. METHODOLOGY

The study employs a real-world dataset obtained from HDD manufacturing, consisting of a single timeframe of production. The dataset contains 53,451 instances and includes 26 features related to components vendors and 17 features related to production machines. All 43 features are categorical data, and the dataset includes two classes: F (denoting defective instances, which number 7,165) and P (denoting non-defective instances, which number 46,286. It is crucial to highlight that the dataset is

complete, devoid of any missing data. This completeness significantly contributes to ensuring the robustness and reliability of our analysis. The R code used for the analysis has been converted into pseudocode and is available in the appendix section. The dataset used is confidential and cannot be shared publicly.

The experiments were conducted utilizing R-3.6.3 software in Windows 10 operating system. To implement the FS techniques, namely Information Gain, Gain Ratio, Chi-Square, and Symmetrical Uncertainty, FSelector library [33] was utilized. In addition, the DT algorithms, ID3 and CART were applied from rpart library [34], while C4.5 was utilized from the RWeka library [35].

The experiments incorporated five-fold cross-validation, where one fold was allocated as the test data while remaining four served as the training data. The experimental procedure encompassed the following steps:

1) Computation of feature importance for all 43 features utilizing four feature selection techniques.
2) Descending order ranking of feature importance for each feature selection technique.
3) Training and testing of each decision tree algorithm with all features as a baseline.
4) Training and testing each decision tree algorithm with each feature selection technique ranking, commencing with the top-ranked feature until achieving the highest accuracy.

The challenge of imbalanced assembly data was addressed by selecting decision tree algorithms as prediction models. These algorithms are appropriate for imbalanced data because they can capture signals from both classes through recursive partitioning. The original data distribution was preserved and any potential bias or noise was avoided by not applying any resampling or weighting methods to the data [36]. The accuracy of the models was measured by the confusion matrix, which displays the number of correct and incorrect classifications for each class. This metric allows the evaluation of the performance of the models on both defective and non-defective drives, irrespective of the class imbalance.

To evaluate the performance of the models, predictive accuracy, total model-building times, the number of rules, and the total test time were utilized. Predictive accuracy was computed based on the confusion matrix, which is a two-dimensional matrix that compares the predicted class values to the actual class values. The confusion matrix reports the number of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN). Predictive accuracy was determined as the proportion of correctly classified outcomes with the total number of outcomes as shown in (7).

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \tag{7}$$

Total model-building time was computed by dividing it into two cases. The first case involves the baseline, whereby the time was computed by performing five-fold cross-validation to all features until the model-building process was complete. The second case pertained to feature selection techniques, whereby the time was calculated based on the feature importance calculation, training, and testing of each algorithm starting with the top-ranked feature until the highest accuracy is attained.

Total test times (TTT) of the current and proposed processes are calculated by (8) and (9), respectively. Total test time of the current process is determined by multiplying number of instances (Q) by the average test time (T). For a given classifier, the aim is to determine defect probability threshold, α, that minimizes the total test time of the proposed process. Therefore, the following quantities are considered: number of true positives at threshold (QTP_α), number of false positives at threshold α (QTP_α), and short test divider (r) where r ranged from 1 to T. If $r = 2$, the short test time is equal to half of the total test time of normal process.

$$TTT_{Current\ Process} = Q \times T \tag{8}$$

$$\begin{aligned} TTT_{Proposed\ Process,\alpha} = {} & (QTP_\alpha \times \frac{T}{r}) \\ & + (QTP_\alpha \times \frac{(1+r) \times T}{r}) \\ & + ((Q - QTP_\alpha - QFP_\alpha) \times T) \end{aligned} \tag{9}$$

In order to validate the significance of the study's findings, a series of statistical tests were conducted. These tests focused on the accuracy of the models and the total test time of the model with the highest accuracy. The null hypothesis proposed that there was no significant difference between the use of all features and the use of feature selection techniques for each decision tree algorithm. Conversely, the alternative hypothesis suggested that the use of feature selection techniques could enhance the accuracy and decrease the total test time compared to the use of all features.

The study employed a t-test to compare the mean values of the accuracy for each model and the total test time for the model with the highest accuracy, setting the significance level at 0.05. The p-values derived from the t-test were reported to represent the likelihood of observing the results under the null hypothesis. Additionally, the study calculated the 95% confidence intervals for both the accuracy and the total test time, providing a range within which the true population mean is likely to fall with 95% certainty.

The study concluded that if the p-value was less than 0.05 and the confidence interval did not encompass zero, it could be inferred that the use of feature selection techniques resulted in statistically significant improvements over the use of all features.

III. RESULTS AND DISCUSSION

Table 2 presents the accuracy of each decision tree algorithm for every feature selection technique, with the maximum values highlighted in bold. All decision tree algorithms for all feature selection techniques exhibited better than using all features. Specifically, the accuracies of ID3, C4.5, and CART with IG were better than using all features, with improvements of 0.0054, 0.0047, and 0.0095, respectively.

TABLE 2. Comparison of accuracy.

Decision Tree Algorithm	Feature Selection Technique				
	All Features	IG	GR	χ^2	SU
ID3	0.8759	**0.8813**	0.8808	0.8812	0.8810
C4.5	0.8770	**0.8817**	0.8809	0.8816	0.8815
CART	0.8718	**0.8813**	0.8808	0.8811	0.8810

TABLE 3. Comparison of total times for building models (second).

Decision Tree Algorithm	Feature Selection Technique				
	All Features	IG	GR	χ^2	SU
ID3	99.0	**28.6**	62.8	41.6	78.0
C4.5	146.5	**27.9**	57.4	29.7	49.4
CART	101.2	**30.0**	62.3	42.9	78.4

TABLE 4. Comparison of number of rules.

Decision Tree Algorithm	Feature Selection Technique				
	All Features	IG	GR	χ^2	SU
ID3	164	50	**29**	73	71
C4.5	387	**7**	**7**	**7**	8
CART	189	49	**28**	80	79

TABLE 5. Comparison of accuracy for each decision tree algorithm at 95% confidence intervals between all features, IG, GR, χ^2 and SU.

Decision Tree Algorithm	Feature Selection Technique									
	All Features		IG		GR		χ^2		SU	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
ID3	0.8737	0.8782	**0.8793**	**0.8832**	**0.8790**	**0.8826**	**0.8792**	**0.8832**	**0.8789**	**0.8832**
C4.5	0.8755	0.8785	**0.8799**	**0.8836**	**0.8791**	**0.8827**	**0.8797**	**0.8835**	**0.8797**	**0.8833**
CART	0.8695	0.8742	**0.8793**	**0.8832**	**0.8790**	**0.8826**	**0.8791**	**0.8832**	**0.8788**	**0.8832**

Table 3 displays the total times taken to build models, in seconds, for each decision tree algorithm with each feature selection technique, with the minimum values highlighted in bold. All decision tree algorithms for all feature selection techniques took less time to build models than using all features. Moreover, all decision tree algorithms with IG took the least time, with 70.4, 118.6, and 71.2 seconds less than using all features, respectively.

Table 4 illustrates the number of rules for each decision tree algorithm with each feature selection technique, with the minimum values highlighted in bold. The numbers of rules for all decision tree algorithms with feature selection techniques were lower than using all features. Specifically, the number of rules for ID3 and CART with GR were lower than with all features, with 135 and 161. Furthermore, the numbers of rules for C4.5 with IG, GR, and χ^2 were lower than with all features, with 380 equally.

Moreover, Table 5 presents the comparison of accuracy for each algorithm between using all features and each feature selection technique, with the results at 95% confidential intervals. Highlighted in bold are results indicating that using feature selection techniques yielded statistically significant improvements over using all features.

Based on the best model, Table 6 shows the total test times in millions of hours between the current process and the proposed process for each threshold value, with the minimum values highlighted in bold. Both total test times are calculated for $T = 720$ hours and $r = 2$. The result shows that the model with the optimal cut-off between 0.15 and 0.70 gives the minimum total test time.

The failure probability threshold range of 0.15 to 0.70 was chosen based on accuracy-efficiency trade-offs, empirical evidence from our experiment, and practical significance for HDD manufacturing. Lower thresholds increase efficiency but also false positives, while higher thresholds decrease false positives but increase testing time. Our experiment showed the optimal range to be between 0.15 and 0.70. This range ensures hard drives with low or high failure probabilities

TABLE 6. Comparison of the total test time (in millions of hours) of the current process and the proposed process for each failure probability cut-off (threshold).

Cut-off	Current Process	Proposed Process					
		Average	Fold #1	Fold #2	Fold #3	Fold #4	Fold #5
0	7.6968	10.5136	10.5044	10.5192	10.4976	10.5098	10.5372
0.05	7.6968	10.5136	10.5044	10.5192	10.4976	10.5098	10.5372
0.1	7.6968	10.5136	10.5044	10.5192	10.4976	10.5098	10.5372
0.15	7.6968	**7.6363**	**7.6410**	**7.6352**	**7.6324**	**7.6367**	**7.6360**
0.2	7.6968	**7.6363**	**7.6410**	**7.6352**	**7.6324**	**7.6367**	**7.6360**
0.25	7.6968	**7.6363**	**7.6410**	**7.6352**	**7.6324**	**7.6367**	**7.6360**
0.3	7.6968	**7.6363**	**7.6410**	**7.6352**	**7.6324**	**7.6367**	**7.6360**
0.35	7.6968	**7.6363**	**7.6410**	**7.6352**	**7.6324**	**7.6367**	**7.6360**
0.4	7.6968	**7.6363**	**7.6410**	**7.6352**	**7.6324**	**7.6367**	**7.6360**
0.45	7.6968	**7.6363**	**7.6410**	**7.6352**	**7.6324**	**7.6367**	**7.6360**
0.5	7.6968	**7.6363**	**7.6410**	**7.6352**	**7.6324**	**7.6367**	**7.6360**
0.55	7.6968	**7.6363**	**7.6410**	**7.6352**	**7.6324**	**7.6367**	**7.6360**
0.6	7.6968	**7.6363**	**7.6410**	**7.6352**	**7.6324**	**7.6367**	**7.6360**
0.65	7.6968	**7.6363**	**7.6410**	**7.6352**	**7.6324**	**7.6367**	**7.6360**
0.7	7.6968	**7.6363**	**7.6410**	**7.6352**	**7.6324**	**7.6367**	**7.6360**
0.75	7.6968	7.6364	**7.6410**	**7.6352**	**7.6324**	**7.6367**	7.6367
0.8	7.6968	7.6368	7.6414	7.636	7.6331	**7.6367**	7.6367
0.85	7.6968	7.6433	7.6734	7.636	7.6331	7.6374	7.6367
0.9	7.6968	7.6692	7.6734	7.6669	7.6694	7.6698	7.6666
0.95	7.6968	7.6787	7.6817	7.6759	7.6792	7.6792	7.6777
1	7.6968	7.6969	7.6975	7.6968	7.6968	7.6968	7.6968

undergo normal testing, optimizing testing time and resource consumption while maintaining quality.

Figure 5 shows the optimal failure probability cut-off versus the total test time for each fold and the average. The figure illustrates that the optimal cut-off ranges between 0.15 and 0.70, which minimizes the total test time of the proposed process.

Figure 6 shows the boxplots of the total test time of the current process and the proposed process at the best threshold for each fold and the average. The boxplots display the minimum, maximum, median, and quartiles of the total test time for each process and each fold. The figure illustrates that the proposed process has a lower total test time than the current process for all folds and the average. The figure also shows that the proposed process has less variation in the total test time than the current process, indicating that it is more consistent and stable. The figure supports the conclusion that the proposed process can significantly reduce the total test time compared to the current process.

Additionally, Table 7 shows the differences in total test times between the current process and the proposed process at the best threshold. Table 8 shows the average difference between the total test time and the 95% confidence interval calculated from the data in Table 7. The result shows that the total test time of the proposed process is significantly reduced compared to the current process.

TABLE 7. Differences in total test times between the current process and the proposed process at the best threshold.

Fold #	(1) Current Process	(2) Proposed Process	(1) - (2)
1	7.6968	7.6410	0.0558
2	7.6968	7.6352	0.0616
3	7.6968	7.6324	0.0644
4	7.6968	7.6367	0.0601
5	7.6968	7.6360	0.0608
		Mean	0.0605
		Standard deviation	0.0031

TABLE 8. Mean total reduction test time at 95% confidence intervals between normal process and proposed process.

Total reduction test time (Million Hrs.)		
Mean	**Min**	**Max**
0.0605	0.0578	0.0633

A. RULE GENERATION AND APPLICATION

The rules were generated using the C4.5 algorithm with IG as the feature selection technique. This algorithm uses categorical features (C1, C2, …, C43) to classify hard disk drives into passers (P) or defectives (F). Each feature has a set of possible values denoted by V0, V1, V2, etc. The rules

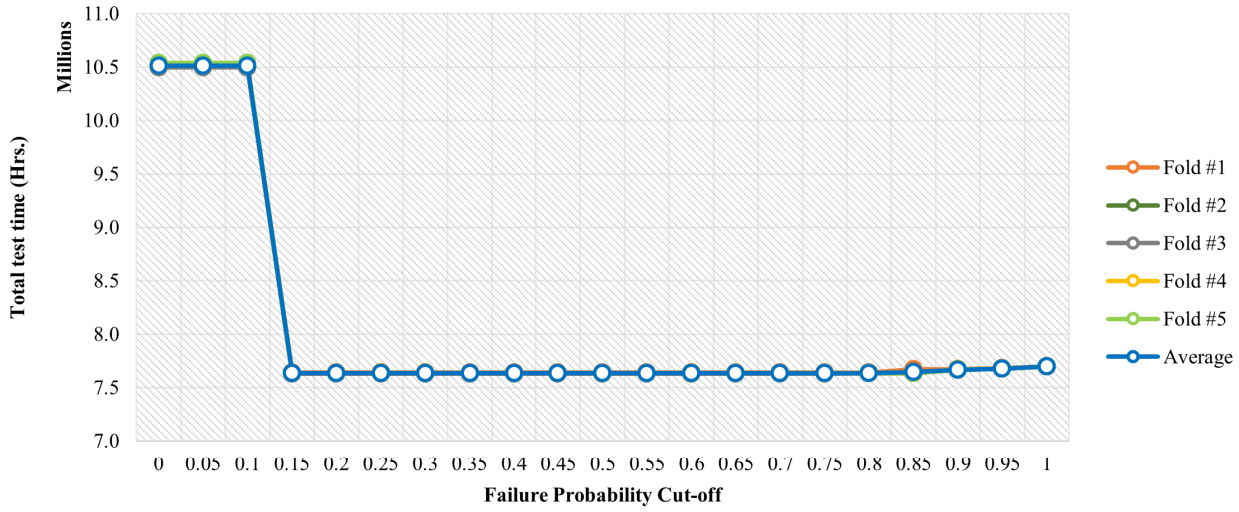

FIGURE 5. A plot of the optimal failure probability cut-off versus the total test time for each fold and the average, showing the range of cut-off values that minimize the total test time of the proposed process.

are applied in a top-down manner, starting from the first rule and moving to the next rule if the condition is not satisfied. The rules can be written as follows:

- If C29 equals "V0", then the class is F.
- If C29 equals "V1" or "V2" and C23 equals "V0", then the class is F.
- If C29 equals "V1" or "V2", C23 equals "V1" or "V2", and C29 equals "V1", then the class is P.
- If C29 equals "V1" or "V2", C23 equals "V1" or "V2", C29 equals "V2", and C43 is in ["V0", "V1", "V2", "V3", "V4", "V5", "V6"], then the class is F.
- If C29 equals "V1" or "V2", C23 equals "V1" or "V2", C29 equals "V2", C43 is in ["V7", "V8", "V9", "V10", "V11", "V12", "V13", "V14", "V15", "V16", "V17", "V18", "V19", "V20", "V21", "V22", "V23", "V24", "V25", "V26", "V27", "V28", "V29"], and C26 equals "V0", then the class is F.
- If C29 equals "V1" or "V2", C23 equals "V1" or "V2", C29 equals "V2", C43 is in ["V7", "V8", "V9", "V10", "V11", "V12", "V13", "V14", "V15", "V16", "V17", "V18", "V19", "V20", "V21", "V22", "V23", "V24", "V25", "V26", "V27", "V28", "V29"], and C26 equals "V1", then the class is P.
- If C29 equals "V1" or "V2", C23 equals "V1" or "V2", C29 equals "V2", C43 is in ["V7", "V8", "V9", "V10", "V11", "V12", "V13", "V14", "V15", "V16", "V17", "V18", "V19", "V20", "V21", "V22", "V23", "V24", "V25", "V26", "V27", "V28", "V29"], and C26 equals "V2" or "V3", then the class is F.

To apply these rules to a new instance, the values of the features C29, C23, C43, and C26 need to be checked, and they are compared with the conditions of each rule. The class of the instance is determined by the first rule that matches the values of the features. In terms of practical application in manufacturing, these rules can be used to predict the failure probability of hard drives. For example, if a rule states that a hard drive with a certain combination of component and assembly machine is likely to fail, this information can be used to adjust the manufacturing process, such as changing the component or assembly machine, to reduce the failure probability.

IV. ETHICAL CONSIDERATIONS

A. DATA PRIVACY AND SECURITY

The expert system relies on assembly data to identify potentially defective drives before they undergo extensive testing. This data contains confidential information about the components, machines, and vendors that are part of the production process. This data is only used within the manufacturer and not disclosed to any external parties.

B. IMPACT ON WORKERS

The expert system does not affect the workers involved in the production process. Workers do not need to change their roles or tasks. The expert system only assists them in streamlining defect detection and reducing testing time and resource consumption.

C. ENVIRONMENTAL IMPACT

The expert system may have a positive environmental impact by optimizing the manufacturing process of hard disk drives. This may lead to lower energy consumption and waste generation, which could benefit the environment. The expert system aims to achieve high accuracy in defect prediction, which could further reduce the test time.

D. TRANSPARENCY AND ACCOUNTABILITY

The expert system makes decisions based on the rules derived from the prediction model. These rules are created from DT,

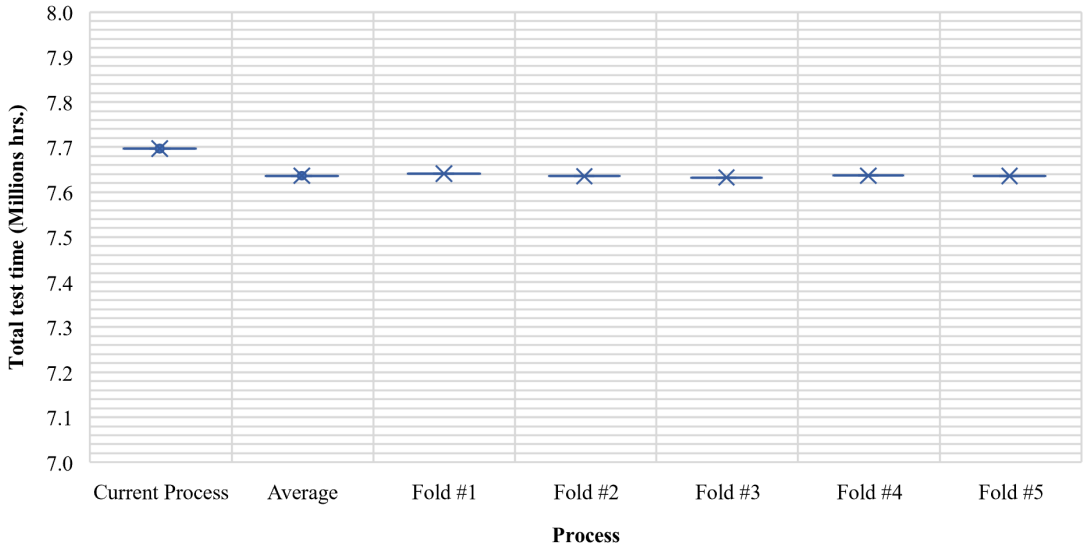

FIGURE 6. Boxplots of the total test time of the current process and the proposed process at the best threshold for each fold and the average, showing the reduction and variation of the total test time of the proposed process.

which are easy to understand and explain by the workers, customers, or regulators. The expert system also verifies its decisions with real test operations (short and full test operations). Therefore, the system poses no risk for the manufacturer.

V. CONCLUSION

This paper aimed to develop a rule-based expert system for enhancing the hard disk drive manufacturing process. The system predicts defective drives before undergoing extensive testing, using feature selection techniques and decision tree algorithms. The paper presented the following aspects:

Summary: The paper proposed a novel rule-based expert system that leveraged assembly data to predict defects and reduce testing time in the hard disk drive manufacturing process. The paper compared various feature selection techniques and decision tree algorithms for building the predictive model, and showed that information gain using C4.5 outperformed the others. The paper also determined the optimal failure probability threshold that minimized the total test time of the proposed process, and demonstrated that the proposed process significantly reduced the test time compared to the current process.

Major Findings: The experimental results revealed that information gain using C4.5 achieved the highest accuracy of 0.8817, required only 27.9 seconds to construct the model, and generated only 7 rules for classification. The results also indicated that the optimal failure probability threshold ranged from 0.15 to 0.70, and that the proposed process with this classifier and threshold reduced the total test time by 0.0605 million hours on average, with 95% confidence.

Contributions: The paper contributed to the literature on hard disk drive manufacturing optimization by proposing a novel rule-based expert system that leveraged assembly data to predict defects and reduce testing time. The paper also contributed to the literature on feature selection and decision tree algorithms by evaluating different combinations of techniques on a real-world dataset. The paper provided a practical and effective solution to a real problem faced by hard disk drive manufacturers, and validated its feasibility and benefits with empirical evidence.

Practical Implications: The proposed approach could improve the efficiency and cost-effectiveness of the hard disk drive manufacturing process by reducing the time and resources required for testing and increasing the capacity of the test equipment. The proposed approach could also enhance the quality and reliability of the hard disk drives by identifying defects early and avoiding unnecessary testing.

Ethical Implications: The proposed approach also raised some ethical implications that need to be considered and addressed by the researchers and the manufacturers. For example, the impact of the predictive model on the workers' well-being and job satisfaction, as the model could potentially replace some of their tasks or change their roles and responsibilities. The environmental and social effects of the increased production and consumption of hard disk drives, as the model could stimulate the demand and supply of the product, which could have negative consequences on the natural resources, energy consumption, waste generation, and digital divide. The ethical responsibility of the researchers and the manufacturers to ensure the transparency, fairness, and accountability of the predictive model, as the model could have errors, biases, or uncertainties that could affect the decision-making and the outcomes of the manufacturing process. The protection of the data privacy and security of the assembly process and the hard disk drives, as the model could expose sensitive or confidential information about the

components, machines, vendors, or customers, which could be vulnerable to unauthorized access, misuse, or breach.

Future Work: Future work could focus on applying the model to other time frames and hard disk drive products, exploring additional feature selection techniques and machine learning algorithms, and further optimizing the approach for greater accuracy and efficiency. Moreover, future work could also address the ethical issues raised by this paper, and propose ways to mitigate the potential risks and enhance the benefits of the proposed approach for the stakeholders involved. Future work could also involve conducting a stakeholder analysis and an ethical impact assessment to identify and evaluate the ethical values, principles, and dilemmas associated with the proposed approach, and to develop ethical guidelines, policies, and best practices for its responsible and sustainable implementation.

APPENDIX. PSEUDOCODE FOR R CODE

Algorithm 1 Library Setup

```
{Define the list of packages}
list_packages <- ["dplyr", "tidyr", ..., "reservr"]
{Install and load packages}
installed_packages <- Get installed packages
for each package in list_packages do
    if package is not in installed_packages then
        Install package from 'https://cloud.r-project.org/'
    end if
    Load package
end for
```

Algorithm 2 Data Preparation

```
{Specify data path}
path <- 'C:/'
{Import data as characters}
dt_raw <- Read CSV file from path
{Data preprocessing section}
col_str <- ""
col_too_many <- ""
col_singleton <- ""
for each column in dt_raw do
    if column has only one unique value then
        Add column to col_singleton
    else if column has more than 100 unique values then
        Add column to col_too_many
    else
        Add column to col_str
    end if
end for
{Select non-singleton columns}
selected_columns <- Split col_str by ","
dt_select <- Select columns "STATUS" and selected_columns from dt_raw
Remove dt_raw
```

Algorithm 3 Feature Selection

```
{Feature Importance section}
feature_importance_algorithms <- ["information_gain", "gain_ratio", "chi_squared", "symmetrical_uncertainty"]

feature_importance_results <- Empty list
for each algorithm in feature_importance_algorithms do
    current_dr <- Calculate feature importance using algorithm on dt_select
    current_dr <- Order current_dr by attribute importance
    Add current_dr to feature_importance_results
end for
```

Algorithm 4 Decision Tree Learning

```
{Feature Selection section}
top_n_values <- Sequence from 1 to 10
for each algorithm in ["ID3", "C4.5", "CART"] do
    for each dr_list in feature_importance_results do
        for each top_n in top_n_values do
            dt_a <- Order dr_list by attribute importance
            dt_sel_recode <- dt_select
            for each val_list in Sequence from 1 to 10 do
                top_n <- val_list
                lst_col <- Select top_n columns from dt_a
                dt_tmp0 <- Select columns "STATUS" and lst_col from dt_sel_recode
                dt_tmp0 <- Shuffle dt_tmp0
                for each K in Sequence from 1 to 5 do
                    df_train <- Select first K rows from dt_tmp0
                    df_test <- Select rows from K+1 to end from dt_tmp0
                    model <- Build decision tree model using algorithm on df_train
                    Evaluate model on df_test
                    Record results
                end for
            end for
        end for
    end for
end for
```

REFERENCES

[1] N. Samattapapong and N. Afzulpurkar, "A production throughput forecasting system in an automated hard disk drive test operation using GRNN," *J. Ind. Eng. Manage.*, vol. 9, no. 2, p. 330, Apr. 2016.

[2] S. Sankar, M. Shaw, K. Vaid, and S. Gurumurthi, "Datacenter scale evaluation of the impact of temperature on hard disk drive failures," *ACM Trans. Storage*, vol. 9, no. 2, pp. 1–24, Jul. 2013.

[3] W. Song, A. Ovcharenko, B. Knigge, M. Yang, and F. E. Talke, "Effect of contact conditions during thermo-mechanical contact between a thermal flying height control slider and a disk asperity," *Tribology Int.*, vol. 55, pp. 100–107, Nov. 2012.

[4] Z.-S. Ye, M. Xie, and L.-C. Tang, "Reliability evaluation of hard disk drive failures based on counting processes," *Rel. Eng. Syst. Saf.*, vol. 109, pp. 110–118, Jan. 2013.

[5] H. Sharma and S. Kumar, "A survey on decision tree algorithms of classification in data mining," *J. Sci. Res.*, vol. 5, no. 4, pp. 2094–2097, 2016.

[6] D. Lavanya and D. Rani, "Evaluation of decision tree classifiers on tumor datasets," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 2, pp. 418–423, Jan. 2013.
[7] A. Jovic, K. Brkic, and N. Bogunovic, "A review of feature selection methods with applications," in *Proc. Inf. Commun. Technol. Electron. Microelectron. (MIPRO)*, vol. 38, Jul. 2015, pp. 1200–1205.
[8] A. Hirunyawanakul, N. Kerdprasop, and K. Kerdprasop, "Efficient machine learning methods for hard disk drive yield prediction improvement," *Int. J. Mach. Learn. Comput.*, vol. 10, no. 2, pp. 240–246, Feb. 2020.
[9] Raimond Spekking / CC BY-SA 4.0 (via Wikimedia Commons). *Weitere Beschreibung Hinzufügen. Von Komponente Des Festplattenlaufwerks*. [Online]. Available: https://commons.wikimedia.org/wiki/File:Seagate_Barracuda_Green_ST2000DL003_-_platter_and_head-0348.jpg
[10] A. Al Mamun, G. Guo, and C. Bi, *Hard Disk Drive: Mechatronics and Control*. Boca Raton, FL, USA: CRC Press, 2017.
[11] M. Taktak-Meziou, A. Chemori, J. Ghommam, and N. Derbel, "Mechatronics of hard disk drives: RISE feedback track following control of a R/W head," in *Proc. Mechatronics, Princ., Technol. Appl.*, 2015, pp. 1–25.
[12] M. Kondo, S. Lim, T. Koita, T. Namihira, and C. Tokoro, "Application of electrical pulsed discharge to metal layer exfoliation from glass substrate of hard-disk platter," *Results Eng.*, vol. 12, Dec. 2021, Art. no. 100306.
[13] P. Pimpanont and P. Chutima, "Application of value engineering in head stack assembly process: A case study," *Int. J. Mater., Mech. Manuf.*, vol. 4, no. 1, pp. 46–51, 2015.
[14] R. Oboe, F. Marcassa, P. Capretta, and F. Chrappan Soldavini, "Realization of a hard disk drive head servo-positioning system with a voltage-driven voice-coil motor," *Microsyst. Technol.*, vol. 9, no. 4, pp. 271–281, Mar. 2003.
[15] H. Shamkhalichenar, C. J. Bueche, and J.-W. Choi, "Printed circuit board (PCB) technology for electrochemical sensors and sensing platforms," *Biosensors*, vol. 10, no. 11, p. 159, Oct. 2020.
[16] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.
[17] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann, 1993.
[18] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA, USA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.
[19] K. Haricha, A. Khiat, Y. Issaoui, A. Bahnasse, and H. Ouajji, "Recent technological progress to empower smart manufacturing: Review and potential guidelines," *IEEE Access*, vol. 11, pp. 77929–77951, 2023.
[20] V. Bondarenko, "Subjective-probability approach to design an expert system for assessment of states of complex systems in conditions of non-regular destructive influences," in *Proc. IEEE Int. Conf. Adv. Trends Inf. Theory (ATIT)*, Dec. 2019, pp. 183–186.
[21] L. Bi, W. Cao, W. Hu, and M. Wu, "A dynamic-attention-based heuristic fuzzy expert system for the tuning of microwave cavity filters," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 9, pp. 3695–3707, Sep. 2022.
[22] F. A. Setiawan, R. Puspasari, L. P. Manik, Z. Akbar, Y. A. Kartika, I. A. Satya, D. R. Saleh, A. Indrawati, K. Suzuki, H. Albasri, and M. Wada, "Ontology-assisted expert system for algae identification with certainty factors," *IEEE Access*, vol. 9, pp. 147665–147677, 2021.
[23] R. Kaur, M. Jain, R. M. McAdams, Y. Sun, S. Gupta, R. Mutharaju, S. J. Cho, S. Saluja, J. P. Palma, A. Kaur, and H. Singh, "An ontology and rule-based clinical decision support system for personalized nutrition recommendations in the neonatal intensive care unit," *IEEE Access*, vol. 11, pp. 142433–142446, 2023.
[24] F. Arévalo, C. Tito, M. R. Diprasetya, and A. Schwung, "Fault detection assessment using an extended FMEA and a rule-based expert system," in *Proc. IEEE 17th Int. Conf. Ind. Informat. (INDIN)*, vol. 1, Helsinki, Finland, Jul. 2019, pp. 740–745.
[25] B. R. Tibbitts, "Flexible simulation of a complex semiconductor manufacturing line using a rule-based system," *IBM J. Res. Develop.*, vol. 37, no. 4, pp. 507–522, Jul. 1993.
[26] J. Li, X. Ji, Y. Jia, and B. Zhu, "Hard drive failure prediction using classification and regression trees," in *Proc. IEEE/IFIP Int. Conf. Dependable Syst. Netw.*, 2014, pp. 383–394.
[27] C. Xu, G. Wang, X. Liu, D. Guo, and T.-Y. Liu, "Health status assessment and failure prediction for hard drives with recurrent neural networks," *IEEE Trans. Comput.*, vol. 65, no. 11, pp. 3502–3508, Nov. 2016.
[28] N. Aussel, S. Jaulin, G. Gandon, Y. Petetin, E. Fazli, and S. Chabridon, "Predictive models of hard drive failures based on operational data," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2017, pp. 619–625.
[29] J. Xiao, "Disk failure prediction in data centers via online learning," in *Proc. 47th Int. Conf. Parallel Process.*, 2018, pp. 1–10.
[30] W. Ahmad, S. A. Khan, C. H. Kim, and J.-M. Kim, "Feature selection for improving failure detection in hard disk drives using a genetic algorithm and significance scores," *Appl. Sci.*, vol. 10, no. 9, p. 3200, May 2020.
[31] A. Siltepravet, S. Sinthupinyo, and P. Chongstitvatana, "Improving quality of products in hard drive manufacturing by decision tree technique," *Int. J. Comput. Sci. Issues*, vol. 9, no. 3, pp. 1–5, 2012.
[32] A. Hirunyawanakul, N. Kaoungku, N. Kerdprasop, and K. Kerdprasop, "Feature selection to improve performance of yield prediction in hard disk drive manufacturing," *Int. J. Electr. Electron. Eng. Telecommun.*, vol. 9, no. 6, pp. 420–428, 2020.
[33] P. Romanski, L. Kotthoff, and P. Schratz. *FSelector: Selecting Attributes*. [Online]. Available: https://cran.r-project.org/web/packages/FSelector/index.html
[34] T. Therneau, B. Atkinson, and B. Ripley. *Rpart: Recursive Partitioning and Regression Trees*. [Online]. Available: https://cran.r-project.org/web/packages/rpart/index.html
[35] K. Hornik, C. Buchta, T. Hothorn, A. Karatzoglou, D. Meyer, and A. Zeileis. *RWeka: R/Weka Interface*. [Online]. Available: https://cran.r-project.org/web/packages/RWeka/index.html
[36] M. S. Kraiem, F. Sánchez-Hernández, and M. N. Moreno-García, "Selecting the suitable resampling strategy for imbalanced data classification regarding dataset properties. An approach based on association models," *Appl. Sci.*, vol. 11, no. 18, p. 8546, Sep. 2021, doi: 10.3390/app11188546.

SUPPAKRIT KIRDPONPATTARA received the B.Eng. degree in computer engineering from the Prince of Songkla University, Thailand, and the M.S. degree in computer science from the King Mongkut's Institute of Technology Ladkrabang, Thailand, where he is currently pursuing the Ph.D. degree in robotics and artificial intelligence (RAI).

PITIKHATE SOORAKSA received the B.Ed. (Hons.) and M.Sc. degrees in physics from Srinakharinwirot University, Thailand, the M.S. degree in electrical engineering from The George Washington University, USA, in 1992, and the Ph.D. degree in electrical engineering from the University of Houston, USA, in 1996. He is currently a Professor with the School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Thailand. His research interests include cyber-physical applications and rapid prototypes in robotics and AI.

VEERA BOONJING received the B.Sc. degree in mathematics from Ramkhamhaeng University, Thailand, the M.S. degree in computer science from Chulalongkorn University, Thailand, and the Ph.D. degree in decision sciences and engineering systems from Rensselaer Polytechnic Institute, USA. He is currently an Associate Professor of information technology with the Department of Computer Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Thailand.