An in Depth Exploration of Machine Learning and Statistical Learning Techniques in Data Analytics

Shangguan Jiang
Nanjing University, Gulou, Nanjing, Jiangsu, China, 210093.
jiangguan23@hotmail.com

Correspondence should be addressed to Shangguan Jiang : jiangguan23@hotmail.com

Article Info
Journal of Computational Intelligence in Materials Science (https://anapub.co.ke/journals/jcims/jcims.html)
Doi: https://doi.org/10.53759/832X/JCIMS202402013.
Received 10 September 2024; Revised from 26 October 2024; Accepted 12 November 2024.
Available online 23 November 2024.

Abstract – In this article, we describe a data analytics methodology for gleaning insights from the production lines of a power transfer unit, such as the critical measurements needed to construct a shim used to align shafts. This study also outlines the most effective methods and analytical methodologies for domains used in determining which measurements are afflicted by faults; determining which measurements are afflicted by shim dimensions; determining which relationships exist between station codes; forecasting shim dimensions; determining which duplicate samples are present in faulty data; and determining which error distributions are afflicted by measurement. Both statistical analysis and analysis based on machine learning (ML) are used to these domains. These findings demonstrate the reproduction rate of defective units, the relative significance of measurement in relation to the shim dimensions, error distribution and faulty units of measurements. The 'PTU housing measurement' was shown to be the most critical measurement out of all the shim dimensions by both statistical and ML-based analyses.

Keywords – Power Transfer Units, Data Analytics, Big Data, Machine Learning, Artificial Intelligence, Housing Measurement.

I. INTRODUCTION

Sensors may generate big data on goods, designs, and materials; nevertheless, it is crucial to make use of accurate information for the proper reasons. To reduce defective goods and get insight into the manufacturing process, manufacturers of power transfer units are required to conduct a thorough evaluation of data retrieved from several sensors in assembly and production lines. Also, manufacturers must choose the most appropriate methodologies while picking analytical methodologies. Nowadays, with the proliferation of high-tech sensors made possible by the Internet of Things (IoT), cyber physical systems (CPSs) capture a massive quantity of information, or "big data". Nonetheless, just a fraction of the data that exists is utilized at now, and much of that data is never put to any use at all. Data analytics reliant on real-time and historical data for fault prediction, estimation of production cost, fault detection, and other purposes allows smart manufacturing. By utilizing big data and analytics, regular industrial maintenance may be converted into predictive maintenance. Predicting the health state of a machine using both current and past data allows for continuous health monitoring. Predictive maintenance is one use of ML technology. Manufacturing has been given a new lease of life thanks to data-driven ML methodologies.

Manufacturing involves large-scale fabrication or assembly of raw materials into final goods. According to Han, Zhou, Liang, Li, and Zhu [1], it is a crucial sector of the global economy, contributing around 16% of GDP in 2019 and yielding a total worldwide production of $13.9 trillion. The efficient production of a large quantity of high-quality goods while keeping costs down is a crucial industrial objective. Nonetheless, if a company lacks the resources and equipment necessary to create and manufacture high-quality items, producing such products may be an extremely costly and time-consuming operation. The evolution of manufacturing over the past few of centuries is nothing short of remarkable. Manufacturing companies in the 18th century sought out machinery to replace human labor in production, ushering in the era known as the Industrial Revolution. The Fourth Industrial Revolution or "Industry 4.0" encompasses three technical trends—intelligence, connectivity, and flexible automation—that aim to further computerize industry.

The advent of Industry 4.0 has ushered in a new era of analytics in manufacturing, introducing the era of "Smart Manufacturing." Some of the Smart Manufacturing domain's attributable components are shown in **Fig 1**. The domains represent a method that relies heavily on technology, namely the Internet of Things (IoT) and other linked devices, in order to manufacture products and keep tabs on operations. It seeks to automate industrial processes in order to improve efficiency, supply chain management, sustainability, and the detection of potential system obstacles before they materialize. When machine learning and artificial intelligence (AI) are applied to manufacturing data, businesses may increase the efficiency of both individual assets and the whole product maintenance process (see **Fig 2**).

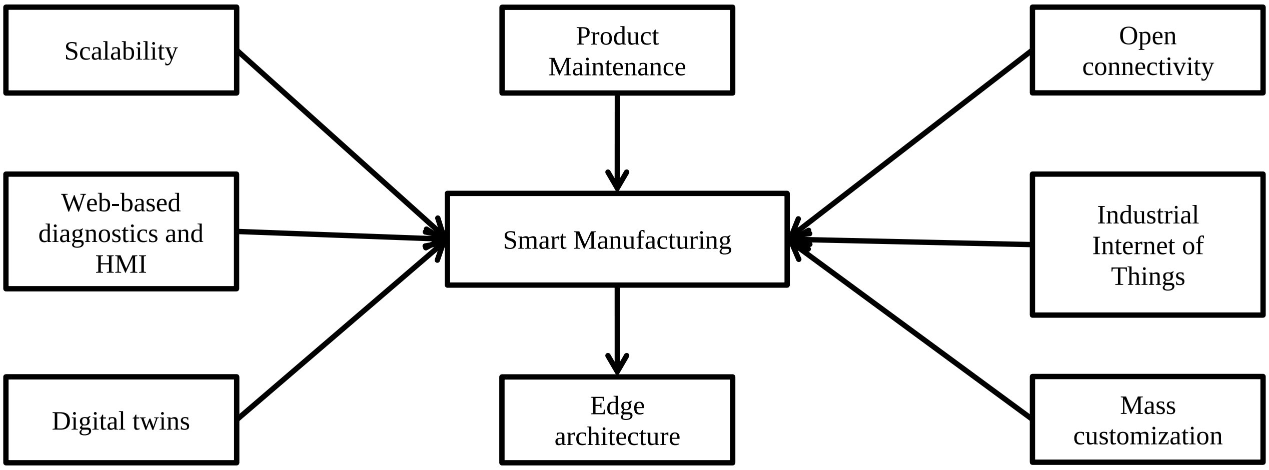

Fig 1. Segments of smart manufacturing

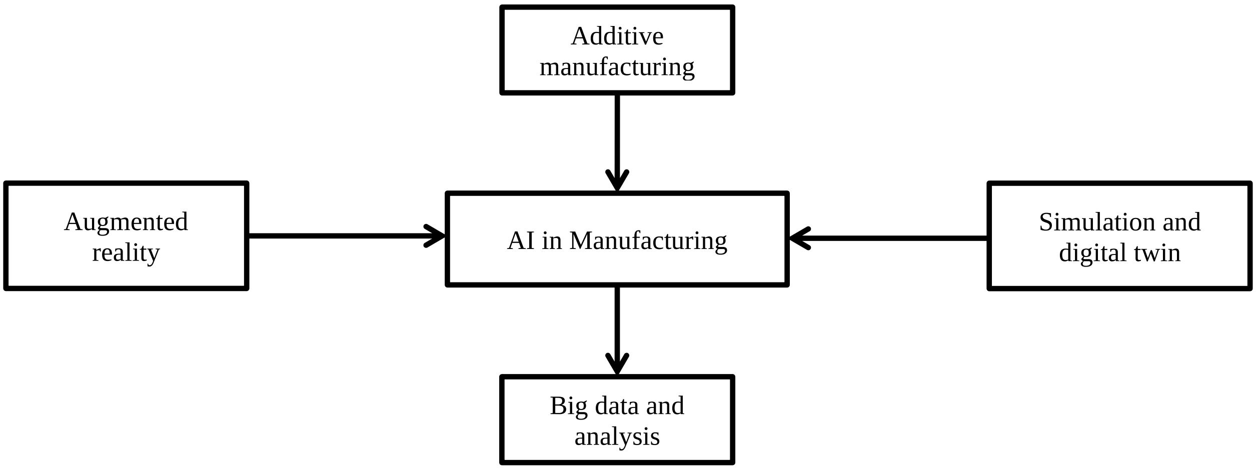

Fig 2. Applications of artificial intelligence and machine learning in the industrial sector

In this paper, we look at how the shim dimensions are affected by various measurement factors (such the 'PTU housing measurement'), statistical analytic techniques (like correlation) and ML algorithms (such as support vector regression). Linear regression and random forest regression) have been used to determine which shim-related metrics are the most important. In addition, the information may be utilized to pinpoint which metrics are most to blame for a malfunctioning device. Prediction of shim dimensions and correlations between station codes are also explored. The error dispersion of assessments and the replication rate of the defective unit are also examined. The optimum methodology to the aforementioned domains is determined by contrasting statistical analysis with ML-based analysis. The rest of the paper is organized as follows: Section II presents an analysis of data collection and analysis of the power transfer unit, dataset, and data analysis. Section III provides an overview of the methodology of data analytics. Results are critically discussed in Section IV, while final remarks are offered in Section V.

II. DATA COLLECTION AND ANALYSIS

Power Transfer Unit

A Power Transfer Unit (PTU) [2] refers to a model, which has the capacity to transfer power from the engine to the drivetrain. To do this, two gears or cogwheels are used. These two gears are crucial to the PTU's operation, and their improper placement causes vibrations and noise. Shims are employed to bring these two gears into proper alignment. Ford Power Transfer Units (PTUs) are now available for replacement at Transtar. The PTU is a kind of All-Wheel-Drive (AWD) [3] transfer case seen in automobiles and SUVs. It may send torque to the rear wheels alone, to the front wheels only, or to both sets of wheels at the same time, depending on the road conditions.

Dataset

Measurements taken on a production line producing PTUs comprise the dataset used in this investigation, which was gathered from the database of a manufacturing firm's logistics in-production network. There are a total of 151,342 built units; however the operator has tagged 6,488 of these as "faulty" owing to measurement discrepancies or improper shim dimensions. The dataset contains forty-two metrics for each unit, such as mounting lengths from the gear heights and gear housing. There is a unique manufacturing date and serial number for each item. The data was gathered at a number of PTU stations, each of which has its own unique station code. In addition, the STATION fields were left blank for good samples and colored red for bad ones.

Data analytics

According to Kulkarni, Kumar, and Rao [4], the phrase "data analytics" is quite general and may be used to refer to a wide variety of methodologies to data analysis. The technique of data analytics could be employed to the different formats of data to provide insights that could be put to good usage. The metrics and trends could be revealed using these methods of data analytics, which might otherwise be witnessed by the massive volumes of data. Based on this understanding, systems and businesses may effectively streamline their daily activities, resulting to improved efficiency. By tracking machines' up- and down-times and the number of jobs waiting to be completed, manufacturers can better allocate resources and get more out of their equipment. There is so much more that can be accomplished with data analytics than just identifying points of contention in the manufacturing process. Companies in the gaming industry use data analytics to create rewards models, which keep the most of the played actively-engaged.

Data analytics are employed by different content designers to keep the players coming back and engaged for more clicks and views. The evaluation of big data is fundamental since it permits businesses to effectively focus on the domains where they can attain the most success. By incorporating it into their daily activities, businesses can find affordable methods do execute business activities and save time by storing big data. Organizational decisions, and customer satisfaction development and trends can be evaluated with the assistance of data analytics, amounting to the enhancement of novel, and developed offerings. **Fig 3** displays the many domains of data analytics that have been explored for this paper.

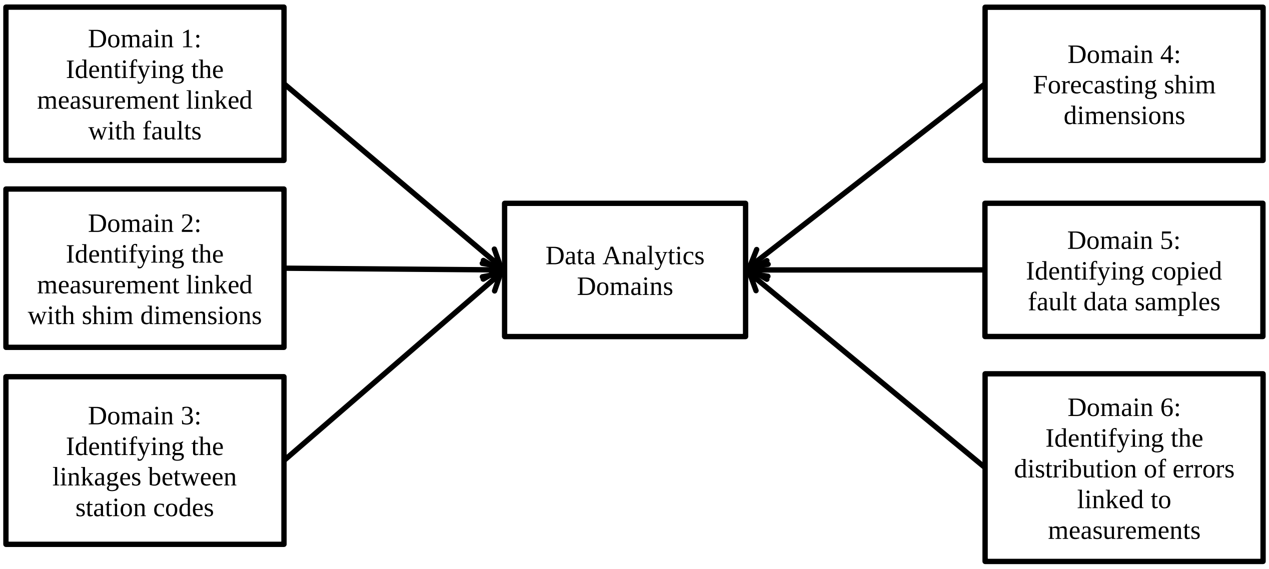
Domain 1: Identifying the measurement linked with faults
Domain 2: Identifying the measurement linked with shim dimensions
Domain 3: Identifying the linkages between station codes
Data Analytics Domains
Domain 4: Forecasting shim dimensions
Domain 5: Identifying copied fault data samples
Domain 6: Identifying the distribution of errors linked to measurements

Fig 3. Several fields of data analytics

Each defective device has a unique station code, which may be used in Domain 3 to establish a connection between them. Dimensions of the shims are predicted in Domain 4, and duplicate samples in the flawed data sets are found in Domain 5.

III. OVERVIEW OF THE METHODOLOGY

Fig 4 depicts this data analytics process in its many stages. Domain expertise, issue definition, data collection and preparation, analytics using traditional statistical methodologies and machine learning-based methods, assessment of the methodology, fresh insights, and the optimal method are all part of the process. A manufacturing firm's assembly line is the first source of domain knowledge, data, needs, and ideas. In most cases, the requirements will dictate how the issue is phrased; here, however, the goal is to learn as much as possible about the assembly line and its processes by posing the appropriate research questions. The results of the method are also evaluated by extracting and storing domain knowledge.

Due to the raw nature of the data, pre-processing was required (including filling in missing numbers and locating any outliers). NaN (not an integer) and null data are converted to zeros and missing values were found and supplied using imputation at this point. In addition, we explored the data to find outliers and inconsistent cardinality. The cardinality of the observations was not one or very low. As a result, the dataset did not include any examples of irregular cardinality. The lowest and maximum values as well as the distribution of the data were studied to spot outliers. Unfortunately, there were no extreme values in the dataset. At the end, we standardized all of the numbers to a scale from zero to one. The set of data was therefore subdivided into training (composed of 80% of data) and tests (comprising 20% of the data) sets so that ML-based analysis could be applied.

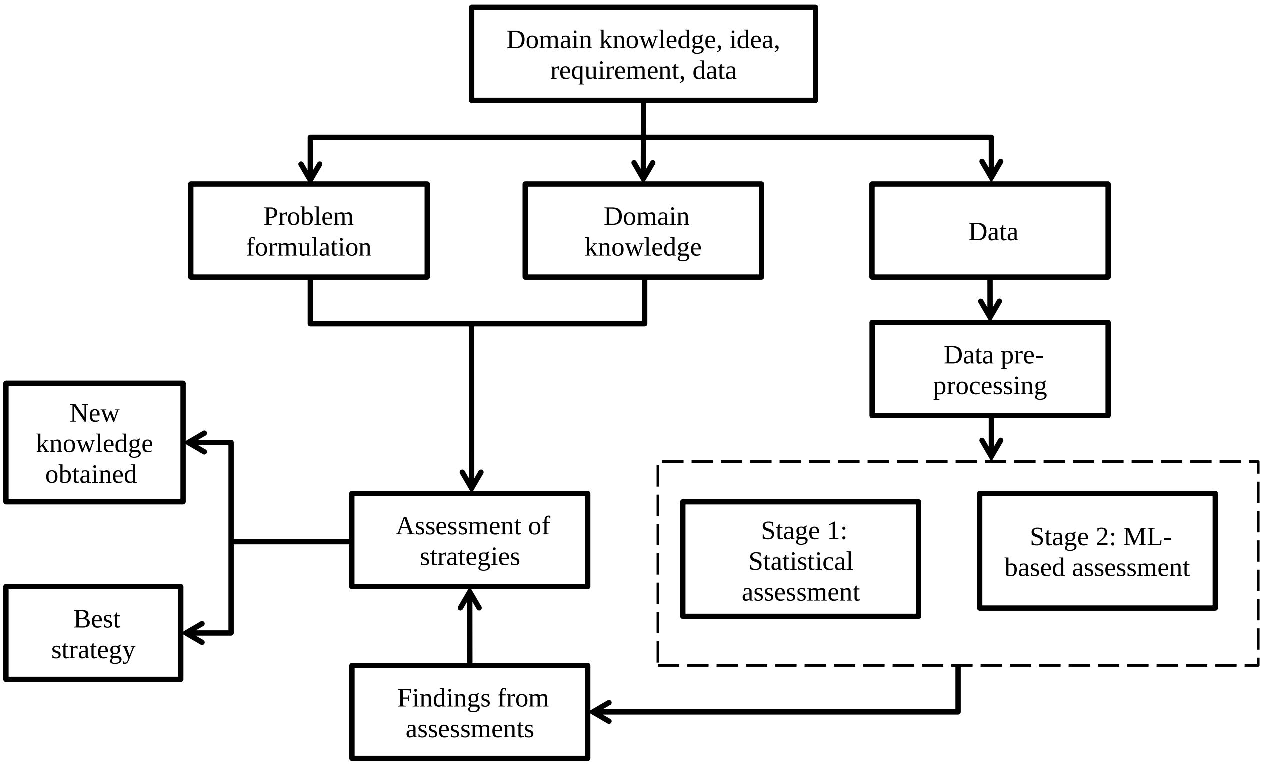
Domain knowledge, idea, requirement, data
Problem formulation
Domain knowledge
Data
Data pre-processing
Stage 1: Statistical assessment
Stage 2: ML-based assessment
Findings from assessments
Assessment of strategies
New knowledge obtained
Best strategy

Fig 4. Several stages of the suggested methodology

In this work, data analytics were carried out in two stages: (1) statistical analysis was carried out in Stage 1 to analyze various data distribution and correlation between various station coding and observations connected with shim parameters to discover linkages of (1) ML-based data analysis and within (2) PTU domains was carried out in Stage 2 to select the most pertinent measurements and minimize the number of observations. The information gained from these two procedures was analyzed and interpreted to provide previously unknown facts about the factory's production process. Furthermore, the best strategies for each domain were determined by contrasting Stage 1 and Stage 2.

Stage 1 of our statistical data analysis was spent investigating and characterizing the dataset. At this stage, we'll look for patterns in the error rate and the location where it was made (the "Assembly Station"), as well as the distribution of defective parts given the various measurement values and correlations between them. Statistical analysis sheds light on the dataset, revealing information such as the relevance of certain measures and the ripple impact of inaccurate readings throughout the assembly line's many stations. The target metrics were binned into 100 categories in order to find the correlations between the various measures and the error rates. The total number of errors for each bin was calculated, and their histogram distribution was examined.

Experts agree that problems with the 'PTU housing measurements,' a crucial metric, lie at the root of the dataset's problems. Faults from distinct stations were associated with the unit code for the 'PTU housing measurements,' applying a correlation analysis in order to determine the dimension to which two randomized observations were geometrically linked. To estimate the correlation, we first compiled a list of 'PTU housing measurements' station codes and then utilized a matrix to determine the cross-correlation of the different codes at the station. The correlation revealed a strong relationship between several stations. It was also discovered that the dataset had many instances of the same flawed samples. Therefore, the erroneous sampling frequency for the measurements was evaluated for every code at the station, and duplicate results matching to a serial number were recognized.

Stage 2 ML-based evaluation sought to categorize PTU faults, forecast shim size, and establish connections between station codes. Fault categorization aids comprehension of the most fundamental measurement, and in the future, it could assist forecasting of different values, which should be altered for a precise unit. Station codes 1 and 0 were assigned to all malfunctioning devices and healthy ones, accordingly. To evaluate how well the ML framework without default and would fare with hyperparameter settings, the models' hyperparameters were fine-tuned. On top of that, most people just used the default settings for hyperparameter tuning, which increased the time it took to build a model by an average of 12 hours. Hyperparameter optimization default settings were not modified due to the lengthy optimization procedure and high performance. For the same reason, no hyperparameters that could be improved were, with the exception of RFR. RFR was tweaked to minimize the gap between the model's anticipated value and the actual value.

The defective devices were classified by training two support vector machine (SVM) classifiers [5] on the dataset. Measurements were ranked using the SVM classifier's coefficient values, and the top-ranked ones were contracted to the experts' recommendations. Both the enhanced and default enhanced hyperparameters were used by one of the classifiers. Box standardized data=0, kernel function='linear, kernel scale=1, and constraint=1are the default hyperparameters for the classifier. Automatic hyperparameter optimization was used in the development of the second classifier. As "auto" was selected for the hyperparameter optimization setting, only "BoxConstraint" and "KernelScale" (two of the available parameters) will be tuned. To facilitate replication, we left all optimization settings at their defaults except for one: "AcquisitionFunctionName," which we changed to "anticipated improvement plus." A model (a support vector classifier) with optimal hyperparameters was developed after 30 iterations. A value of 837.56 for BoxConstraint and a value of 133.58 for KernelScale are optimally possible.

Furthermore, multiple ML techniques (SVR, RFR, and LR) were trained to effectively identity the correlation between PTU housing measurements, Gear (Pinion) height, and Manual adjustments, and to predict shim dimensions. As hyper-parameters were not employed in the input data-point fitting, just a single model was trained using the LR technique. The formula y = bx + c is considered to describe the relationship between input and output. In SVR, we trained two models, one with the hyperparameters tuned, and the second one with default hyperparameters for best performance. Using the default hyperparameters (lambda=8.259106, learner=SVM, regularization=ridge (L2)), SVR learned with the linear kernel. On the other hand, three hyperparameters (BoxConstraint, KernelScale, and Epsilon) were optimized by setting them to "auto" in the optimized model. By default, the optimization option was selected. A regression model with optimal hyperparameters was developed after 30 iterations. KernelScale = 0.013568, Epsilon = 0.00022608, and BoxConstraint = 0.022683, are the optimized hyperparameters.

One of the models was trained using default hyper-parameters, another was trained with four hyperparameters optimized, and a third was trained with all hyperparameters tuned in RFR. These are the hyperparameters used to train the default RFR based on the application of bagged ensemble of 200 distinct trees under regression analysis; Use 200 iterations of ensemble learning, a learn rate of 1, the bag methodology, and all predictors in each split. To optimize the following four hyperparameters in the RFR models, we set them to auto: Method, NumLearningCycles, LearnRate, and MinLeafSize. There weren't any custom optimization settings, thus everything was left at its default. There were 30 iterations used to find the optimal values for four hyperparameters in the RFR model. NumLearning Cycles=85, Method='LS Boost,' and MinLeafSize=1, LearnRate='0.050891 are the optimal hyperparameters. All the fundamental parameters were tuned in the third model. The hyperparameters were tuned to the following values: Method='Bag,' LearnRate=NaN, NumLearning Cycles=16, MinLeaf Size=4, NumVariablesToSample=2 and MaxNum Splits= 60006. Then, the models were put to the test on the test dataset.

About 10 sets of rules were obtained based on the application of the Apriori methodology employed on the Weka platform to determine the connections between the various stations. By changing the value of the 'vehicle' variable to false, class association rules were bypassed in favor of mining general association rules. The rules were ranked using confidence values, with a cutoff of 0.9 for inclusion in the ranking. There was a maximum guarantee of 1.0 in terms of minimum support.

IV. RESULTS AND DISCUSSION

The purpose of this paper was to determine the most effective methods for each step of the assembly process and to get fresh insights into the assembly line as a whole. To determine the optimal ML model, this research makes use of an exploratory validation strategy. **Fig 4** describes and evaluates many subfields of data analytics.

In Domain 1, the manufacturing firm's experts supplied a collection of the most instructive measures that might be mapped to defects. As its name suggests, an expert system is a computer program designed to make judgments and recommendations with the same level of expertise as a human expert. It mimics human behavior by applying human knowledge to situations normally solved by humans. It's a classic case of a system built on information and experience. **Fig 5** is an example of an ES, or expert system. According to **Fig 5**, there are three primary stages to an ES: (1) some kind of knowledge base, (2) some sort of problem-solving and inferential engine, and (3) some sort of human-machine interface.

Hence, an ES is a smart computer program that takes user input through the user interface and makes logical judgments via the inference mechanism based on the information in the knowledge base. Long-term memory (the knowledge base) is where the expertise of the specialists is kept. In the knowledge base, you'll find IF...THEN rules, data, and guidelines. The old adage goes something like this: "Knowledge is power." It is impossible to find solutions to problems without first understanding them in great depth. The role of the knowledge engineer is to write the code required to integrate domain-specific expertise into the ES. Using the data in the knowledge base, the inference engine makes educated guesses (inferences). The software is what allows for the interpretation of commands and the retrieval of relevant information in order to address a specific issue.

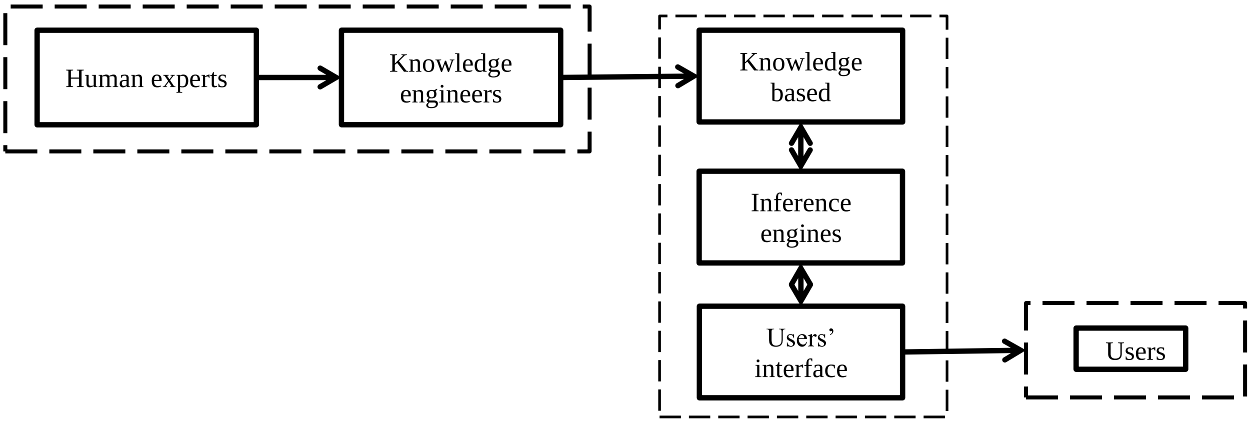

Fig 5. Expert System configuration

The first stage's goal was to evaluate the coefficient of correlation identifiable in every 42 STATION and measurements. Nevertheless, it was discovered that this methodology took a long time. The 4242 matrix generated by the MATLAB program "corrplot" for identifying correlations was not easily interpretable. Analysis of variance (ANOVA) is another way for executing Stage 1 analysis, whereby p-values are utilized to choose the most fundamental measurements. In [6], Lee and Park dispersed data based on a p-value. As the dataset was not generally distributed, the ANOVA technique was not used in this study. Following the steps by Mishra, Jothi, Urolagin, and Irani [7], it is possible to implement Stage 1 analysis. Finding the most crucial metrics required the writers to eliminate them one by one, which was a time-consuming process. These issues prevented us from considering

Stage 1 is an acceptable analytical technique. Stage 2 consisted of identifying a new set of metrics that were important (ML algorithms). One SVM classifier was made using the default hyperparameter values, whereas the other was made with improved hyperparameters. The identified essential measures detected with both the SVM classifiers [8] are illustrated in **Table 1**. Both classifiers generated similar measurements in relation to their significance. However, there was a lot of concordance between the expert measures and the measurements found by the ML method SVM. As a result, SVM classifications were employed to classify the data into either non-faulty or faulty categories. After that, we compared the linear coefficients that went along with our measurements (the predictors). These are the 18 most important indicators that we have compiled. SVM's discovery of additional measures matches the manufacturer-supplied list of 18 measurements. Discussions with subject-matter specialists indicated that whenever an issue arises, service personnel may examine **Table 1**'s measures for signs of malfunction.

Table 1. Essential measurements identified with the SVM aid in Domain 1

Measurements	Description
MECHM3	Lower bearing diameter (cover side)
HUSMATNINGM3	Housing measurement total height
LOCKSHIMSM9	The used shim measurement value
SEKVENS_HISTORIK	History
ADJ3	Adjustment value
CARTRIDGESHIMSM5	Actual shim pinion
PINJONGSHIMSM5	Actual shim measurements value
PINJONGMATNINGM2	Bearing diameter pinion (gear)
MECHM4	Upper bearing diameter (house side)
ADJ2	Adjustment value
LOCKSHIMSM8	Cover shims for leveling the crown wheel shafts within the assembly
MECHM5	Measurements of integrated roundness and centricity on gear sets
HUSMATNINGM5	Bearing seats diameter
HUSSHIMSM4	Calculated house shims
CARTRIDGESHIMSM4	Gear (pinions) shim calculated
PINJONGMATNINGM2	Bearing diameter gears (pinion)
PINJONGKASTM2	Measurements of integrated roundness and companion flanges
HUSSHIMSM6	The variation between used shims and calculated shims

Fig 6 displays the results of classifying the dataset of the tests based on the application of classifiers with default set of hyperparameters and the optimized set of hyperparameters [9]. These metrics demonstrate the classifiers' usefulness. Each classifier received a perfect score on measures of accuracy, specificity, and sensitivity; none of the samples were wrongly labeled as defective or normal. To test for improved efficiency, one reason to develop a hyperparameter-optimized model [10] exists.

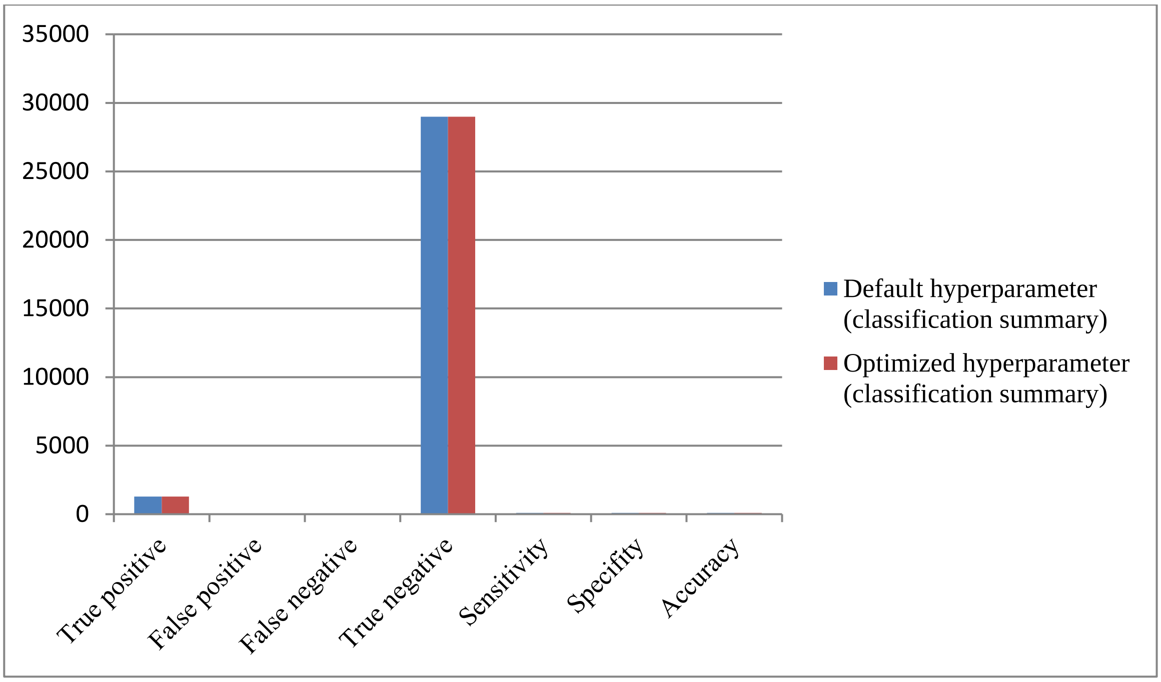

Fig 6. Non-faulty and faulty classifications based on test dataset of SVM in Domain 1

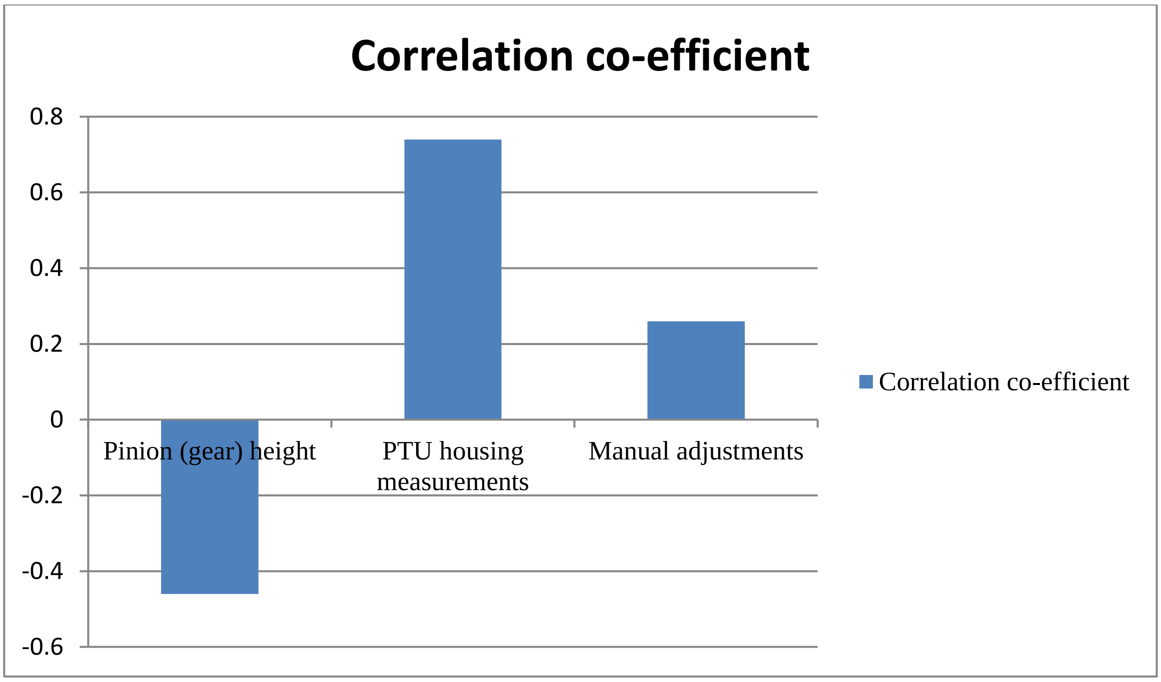

Fig 7. Correlations between the shim dimensions defining measurements in Domain 2

It is also shown that Stage 1 analysis is unfit for use in Domain 1. The complexity of performing Stage 1 grows exponentially as the number of measurements rises. Because of the length and complexity of the implementation process, this domain is better suited for Stage 2.

In Domain 2, the studies were carried out in two Stages: Stage 1 and Stage 2. The shim size was correlated with the 'Gear (Pinion) heights,' 'PTU housing measurements,' and 'Manual adjustments' variables. Correlation coefficients were determined for these data in respect to the shim dimensions in the first Stage, and the findings are indicated in **Fig 7**. **Fig 7** illustrates that 'PTU housing measurements' correlates most strongly with the 'shim dimension,' which is consistent with the consensus of experts. In Stage 2, the LR, RFR, and SVR ML algorithms, with both their basic hyperparameters and their adjusted hyperparameters, detected the relative significance (such as linear measurement coefficients concerned with shim dimensions) (**Fig 8**). These ML techniques made use of regression models to forecast the shim size.

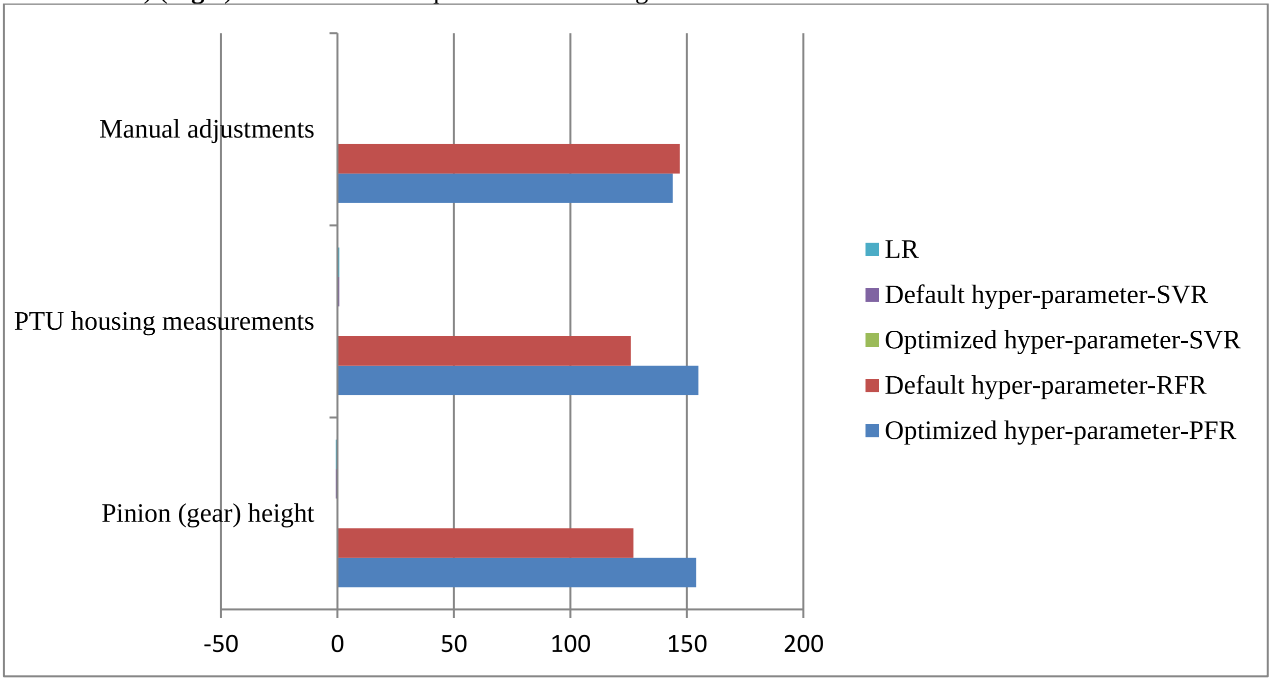

Fig 8. The coefficient of linear regression linked with shim dimensions defining a measurement in Domain 2

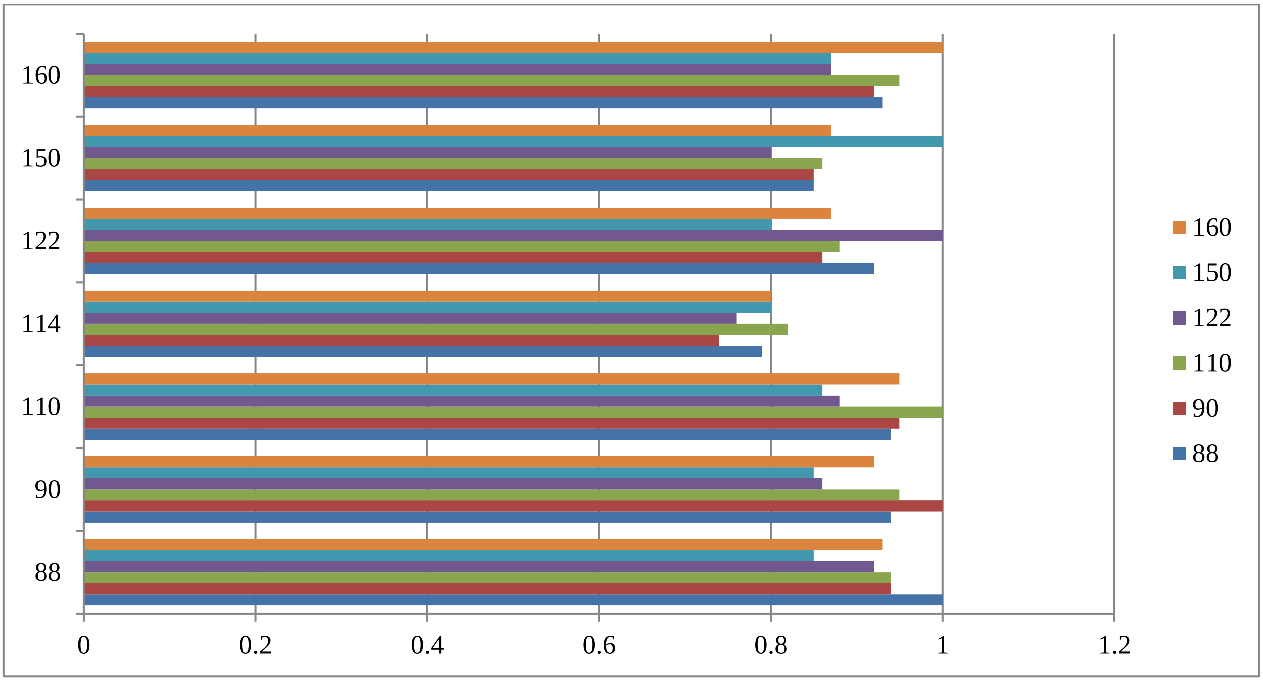

Fig 9. The correlation of the codes within the station in Domain 3

In **Fig 8**, the negative values show a change in measurement in a positive locus compared to shim dimensions that will transform in a negative locus. If the shim size is incorrect, it is likely that the 'PTU housing measurements' is also identified as inaccurate, as illustrated in **Fig 8**. A technician can verify whether or not this measurement has to be tweaked. With the exception of the default hyper-parameters RFR model, each of the other models yielded the same outcome. The 'gear (pinion) height' hyperparameter carries the most weight in the basic RFR model when it comes to shim size. This finding, however, contradicts the predictions of the other theories. Although the hyperparameter-based LR and SVR had better accuracy (**Fig 11**), we prioritized the 'PTU housing measurements. In addition, while comparing the default hyperparameter model with the optimized hyperparameter model, it was found that the optimized model had a lower general relative predictor value. In comparison to the standardized hyper-parameter framework, an optimized hyper-parameter framework minimizes the effects of predictors on shim dimensions.

Nonetheless, even though both the first and the second stage studies were employed in this domain. Stage 1 was more user-friendly. Regression models with hyperparameter adjustment were created in the second Stage. As an added complication, Stage 2 analysis implementation requires familiarity with ML. When a problem's solution can be found using more conventional methods of mathematics or statistics, there's no use in using ML to get there. Thus, Stage 1 analysis is the best choice for this domain. For the 'PTU housing measurement' in Domain 3 we computed the correlation (Stage 1) between different codes of the station and compiled the results in **Fig 9** below (i.e. issues where the coefficient of correlation is more than 0.8). **Fig 9** does not include the codes for the other stations since their correlation coefficients were too low to be meaningful.

In **Fig 9**, the graph “88” shows a coefficient of correlation of the codes “88” with the different stations. **Fig 10** displays the outcomes of association rule mining conducted in Stage 2 analysis utilizing the Weka platform. The confidence in each rule is more than 90%. The first row, for instance, may be interpreted as follows: if Station 114 is fault-free, then there is a one-hundred-percent probability that Station 140 will also be fault-free. If the lift is bigger than 1, the rule heads and rule body appear together often compared to how they would be predicted. Rule bodies and rule head may be considered separate if the certainty value is 1. Rule improvement is indicated by a conviction value greater than 1. More often than not, the rule heads and rule body will occur simultaneously if the leverage value is large. According to **Fig 10**, these metrics all point to the rules being dependable.

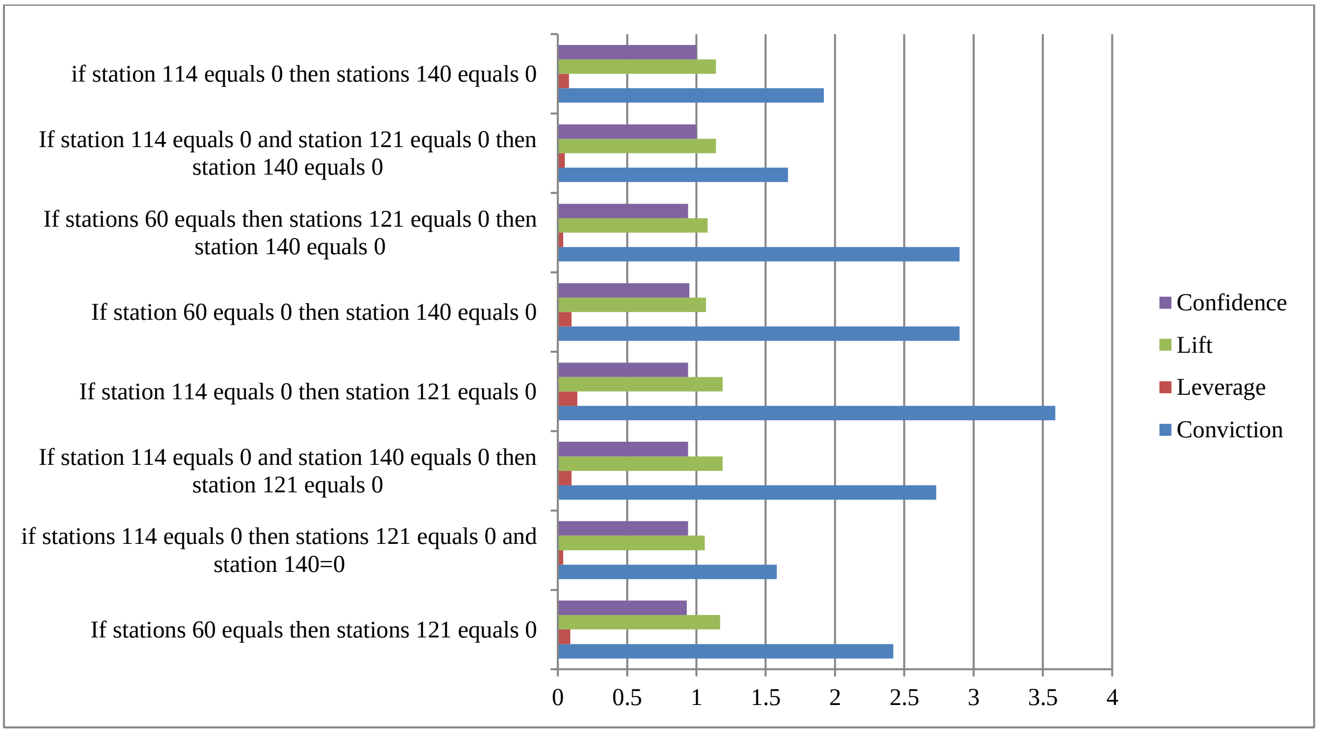

Fig 10. Rule extracted from the codes at the station in Domain 3

Stage 2's highly correlated stations, however, don't match up with Stage 1's findings. Stage 2 is more precise, according to manual checks of the stations. The absence of a defect was disregarded in favor of counting those that did occur in the statistical analysis. ML took into account both fault and non-fault station connections. Thus, Stage 2 analysis is optimal for this field.

Stage 1 in Domain 4 makes use of particular statistical methods chosen after an evaluation of fifty publications published in 2019 and 2020 that have through a rigorous peer review process. We looked at using spatial statistics, for instance, but that methodology is best suited for use in the extraction of features and no predicting. The Cox proportional hazard modeling was also used to forecast when an event will occur, but it was unable to account for the shim dimension. AFT (accelerated failure time) was also taken into account. Nonetheless, the same procedure as the Cox exponential hazards regression is used in this model. Although logistic extrapolation is a classifier and not a regression tool, it was examined in just one research. Consequently, we were unable to identify any alternative statistical methods suitable for use in Domain 4. As a result, Domain 4 did not participate in Stage 1.

Stage 2 results showed that the shim dimension could be predicted with near-perfect precision by both the LR or SVR (standard and modified hyperparameter) methods. As contrasted to LR and SVR, RFR's (both default and customized hyperparameter) projected value was found to be only slightly off from the actual value. One of the RFR simulations had all of its hyperparameters tuned, however that model also had the largest variance. Using the testing dataset and the improved hyperparameter RFR method, a parity graph for the shim-dimension predictions is a deviation from the true numbers was no more than 10%.

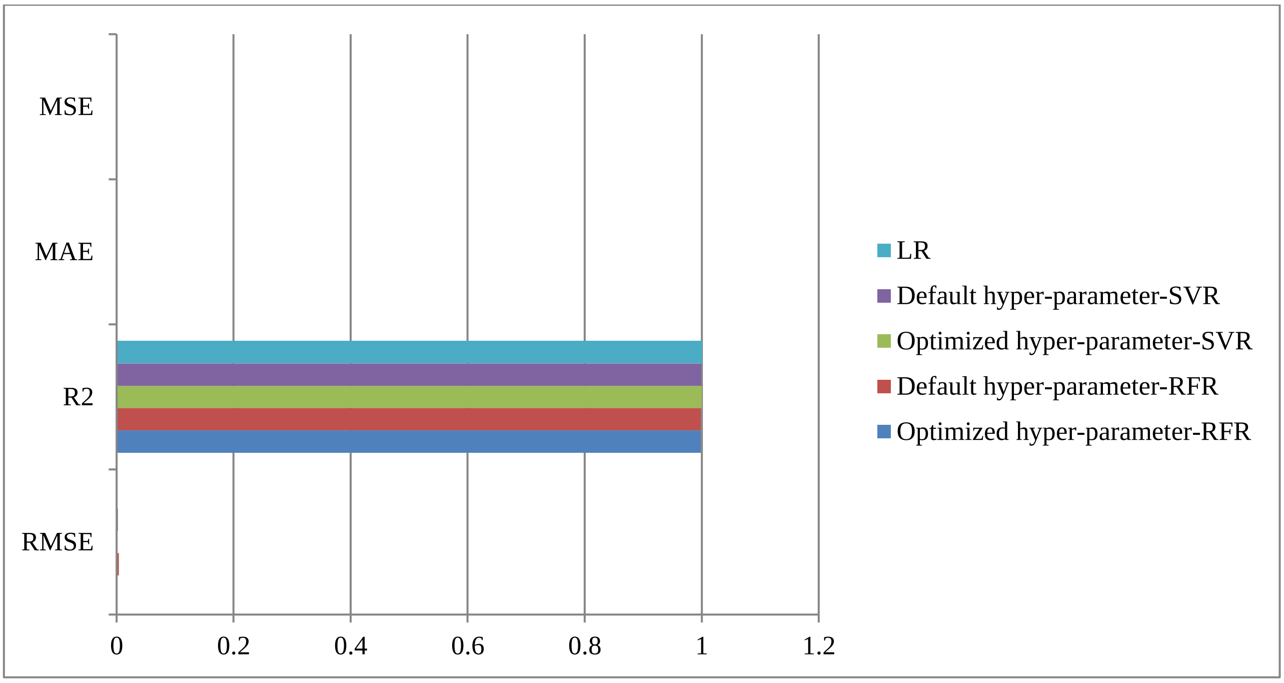

Fig 11. Error rate based on regression frameworks on the dataset in Domain 4

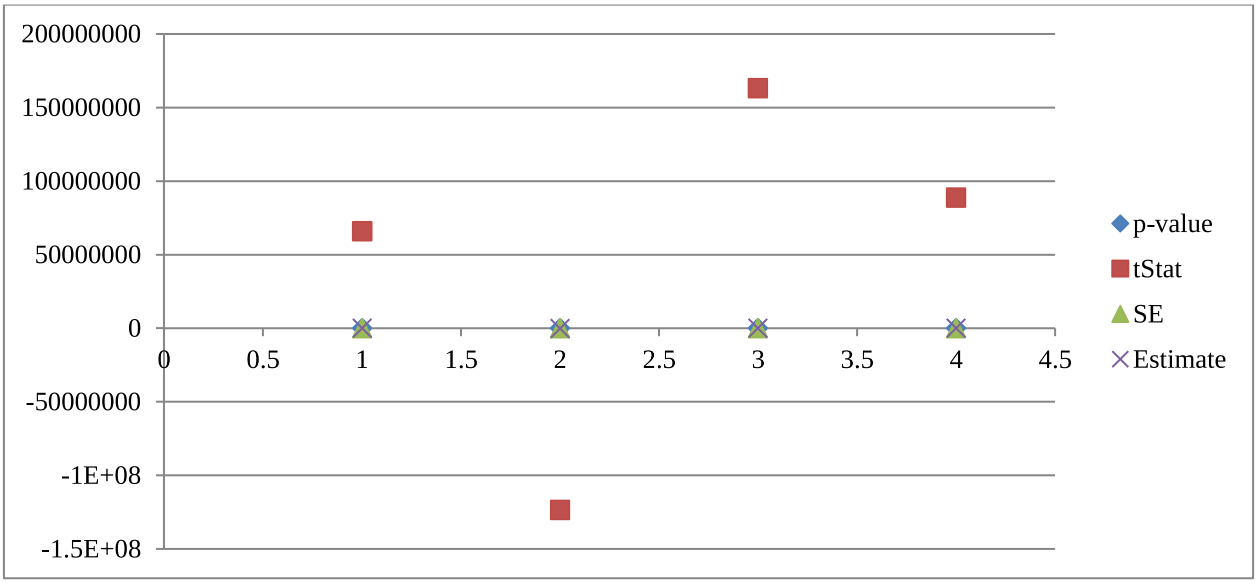

Fig 12. Projected co-efficient of the model of linear regression in Domain 4

Regression results such as R2, MAE, RMSE, and MSE are shown in **Fig 11**. (optimized and default hyperparameter). The optimized frameworks had somewhat better values for R2, RMSE, MAE, and MSE than the default hyperparameter simulations. Nevertheless, the RFR model showed little improvement after hyperparameter optimization. As can be seen in **Fig 11**, the measured data points are quite close to the model's predicted values, and a reduced RMSE value implies a better match. Instead, the models show a high degree of accuracy in predicting the shim size (R2 values of 1 or close to 1)

The models also have very small MAE and MSE values, which indicates that they make accurate predictions. Nevertheless, since technicians labelled the dataset used for comparison, there is a chance that some of the labels are inaccurate. Thus, the model might have certain flaws. The calculated coefficients for the linear regression modeling are shown in **Fig 12** for the variables "Gear (Pinion) height," "PTU housing measurement," and "Manual adjustment." Values of the predictor' coefficients, denoted by the word "Estimate," reveal their relevance in the framework. Of the three, "PTU housing measurement" is the furthermost significant. The SE measures how well the model can estimate coefficient values, and it is denoted by the standard deviation of the estimate, or "SE." Lower SE values represent more accurate estimates. **Fig 12** shows that the model provided a reasonable approximation of the true values of the coefficients due to the tiny SE.

'tStat' evaluates the accuracy of estimate values to decide if a null hypothesis must be rejected or accepted. When the term "null hypothesis" is used, it means that no correlation exists between the two variables. When using a regression model, a higher tStat value indicates a more reliable estimate. Because tStat is high, we may conclude that the null hypothesis is false. P-values in linear regression evaluation indicate whether or not the hypothesis should be ignored. If the p-value is small, we can rule out the null hypothesis in this investigation. The input and the result also have a high degree of association. **Fig 12** shows that there is a perfect correlation between all of the predictors and the outcome, since all of the p-values are 0. Since Stage 1 was not feasible, Stage 2 is the preferable methodology for that particular Domain 4.

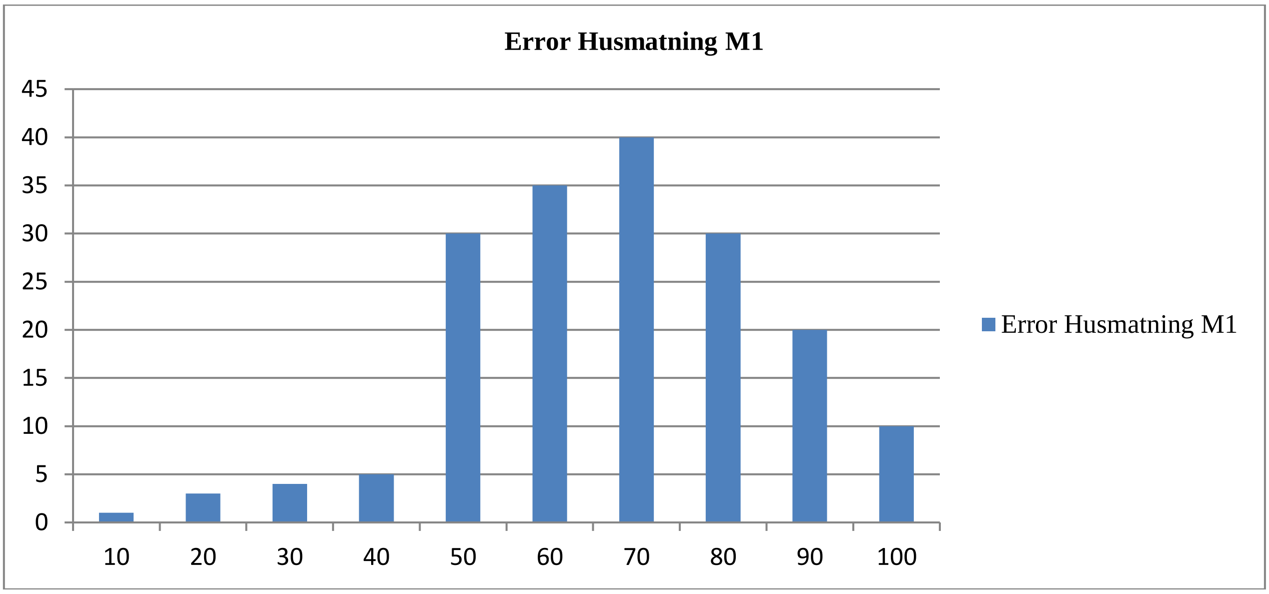

Fig 13. Modeling the 'PTU housing measurements' error distribution

Duplicate PTU units in Domain 5 were identified by looking in the "Serial number" column, and if one already existed, a new one was made in case of a problem. The broken item was fixed and returned with the same "serial number." Station fault codes 90 and 110 were analyzed in Stage 1 of the project. It was discovered that 3,930 products from stations 90 and 110 were defective. Just 360 identically broken pieces were fixed out of a total of 3,930. Experts say that faulty PTUs might be given new "Serial numbers" or scrapped entirely. Although it is not fundamental to employ ML to determine copied cases within a particular set of integers, Stage 2 was not implemented here. The following scenarios call for the use of ML: Complexity, memory requirements, and the need for flexibility are all indicators of an unhumanly difficult endeavor. So, the Stage 1 strategy is the most appropriate for use in Domain 5.

In Domain 6, the first stage was put into action to determine the error rate distribution. There is a Gaussian association between defects and measurements with one exception: "housing measurements from loading residential home," which has a huge bar at the 59%. We take it as given that the equivalence of the data to 59 was not the result of a coding error. These numbers have been verified as accurate after extensive examination. In **Fig 13**, we can see how the 'PTU housing measurement' error is distributed. The error rate is rather high with this value (103.58). On the other side, the rate of errors drops as it falls below 103.68. Due to the identical issues discussed in Domain 5, Stage 2 was never implemented; as a result, Stage 1 is the optimal strategy for this particular domain.

V. CONCLUSION

This research gives a comparison assessment of the appropriateness of different analytical methodologies in the preceding six domains, and adds to our understanding of an assembly line in a manufacturing organization. When a problem is detected in a PTU, the suggested methodologies enable assembly line personnel to just verify the critical measures highlighted by ML (Domain 1) rather than all 42 measurements. A technician may also verify whether the 'PTU housing measurements' fits the shim measurements (Domain 2). The manufacturing firm may learn more about the patterns and root causes of failures by examining the connections between station codes in Domain 3. To aid technicians in making the best shim selection in the event of a mismatch, a cloud-based solution is available for predicting the shim dimensions (Domain 4). Technicians might strive to slow the pace at which defective units are produced by keeping an eye on Domain 5. It has been determined via consultation with manufacturing firm specialists that the 'PTU housing measurements' error distribution (Domain 6) is exponential. While a normal distribution was expected, this investigation discovered a Gaussian one; this disparity will be explored in future studies.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding was received to assist with the preparation of this manuscript.

Competing Interests

There are no competing interests.

References

[1]. Y. Han, L. Zhou, Y. Liang, Z. Li, and Y. Zhu, “Fabrication and properties of silica/mullite porous ceramic by foam-gelcasting process using silicon kerf waste as raw material,” Mater. Chem. Phys., vol. 240, no. 122248, p. 122248, 2020.

[2]. X. Wang, J. Xu, S. Lu, S. Ren, M. Leng, and H. Ma, “Single-receiver multioutput inductive power transfer system with independent regulation and unity power factor,” IEEE Trans. Power Electron., vol. 37, no. 1, pp. 1159–1171, 2022.

[3]. Y. Yang, P. Li, H. Pei, and Y. Zou, “Design of all-wheel-drive power-split hybrid configuration schemes based on hierarchical topology graph theory,” Energy (Oxf.), vol. 242, no. 122944, p. 122944, 2022.

[4]. A. R. Kulkarni, N. Kumar, and K. R. Rao, “Efficacy of Bluetooth-based data collection for road traffic analysis and visualization using big data analytics,” Big Data Min. Anal., vol. 6, no. 2, pp. 139–153, 2023.

[5]. J. Han, T. Zhang, Y. Li, and Z. Liu, “RD-NMSVM: neural mapping support vector machine based on parameter regularization and knowledge distillation,” Int. J. Mach. Learn. Cybern., vol. 13, no. 9, pp. 2785–2798, 2022.

[6]. J. Lee and M. Park, “Estimation of p-values with reduced set of genomic association data,” Korean Data Anal. Soc., vol. 20, no. 4, pp. 1633–1643, 2018.

[7]. R. K. Mishra, J. A. A. Jothi, S. Urolagin, and K. Irani, “Knowledge based topic retrieval for recommendations and tourism promotions,” International Journal of Information Management Data Insights, vol. 3, no. 1, p. 100145, 2023.

[8]. R. Gao and K. Ye, “SVMs-SKSM: Protein Function Multi-label Classification based on SVM-SVM classifiers Fusion and sequences kernel similarity matrix,” Research Square, 2022.

[9]. X. Fan, Y. X. R. Wang, P. Sarkar, and Y. Yue, “A unified framework for tuning hyperparameters in clustering problems,” Stat. Sin., 2024.

[10]. D.-M. Ge, L.-C. Zhao, and M. Esmaeili-Falak, “Estimation of rapid chloride permeability of SCC using hyperparameters optimized random forest models,” J. Sustain. Cem.-based Mater., pp. 1–19, 2022.