among the most popular methods used on Kaggle. Nielsen (2016) compared XGBoost with Gradient Boosting Machine in which shoes that (1) the XGBoost algorithm propose a theoretically justified weighted quantile sketch for efficient proposal calculation, (2) the introduction of a novel sparsity-aware algorithm for parallel tree learning. (3) enabling an effective cache-aware block structure for out-of-core tree learning (T. Chen & Guestrin, 2016). In order to improve the accuracy, XGBoost also adjust some part from original Gradient Boosting Machine: (1) original Gradient Boosting Machine which uses CART (Classification and Regression Trees) as classifier. XGBoost not only can be applied based on CART but also can use linear classifier. (2) XGBoost algorithm apply second order in the Taylor's expansion to loss function, and it also add in regularization, those steps effectively simplify learning models and avoid overfitting. Yet there are some situations that Gradient Boosting Decision Tree (GBDT) can't perfectly predict. For example, when the feature dimension is high and data size is large, the efficiency and scalability are still unsatisfactory. Ke et al. (2017) proposed LightGBM as a highly efficient gradient boosting decision tree to solve to related issues.

2.4 Genetic Algorithm

Inspired by Charles Darwin's theory of natural evolution (Darwin, 2004), nominated survival of fittest concept, Evolutionary algorithms (EAs) are stochastic search methods simulated natural evolutionary process. Evolutionary operations are first proposed in 1950s such as manufacturing plant viewed as evolving species (Box & Hunter, 1957). Being as the core of learning machine in the simulation perspective, EAs are implemented in gradual improvement of a computer programming (Friedberg, 1958). Genetic algorithm (GA), based on the evolutionary strategies of mutation, crossover and selection is proposed by John Holland in the early 1970s, and particularly his book Adaptation in Natural and Artificial Systems (Holland, 1975). Since then GA has become popular and implemented GA in search high quality solutions, optimizing deterministic problems and machine learning domains (Booker, Goldberg, & Holland, 1989). Moreover, the concept of genetic programming that evolution programs will be the combination of Genetic Algorithm and data structure (Michalewicz, 1996) has promote GA as a powerful optimizer for the next decades.

The primary foundation of GA regarding the best known EAs is that GA follows the five basic components (Koza, 1994): First, a genetic representation of potential solutions to the problem is necessary for GA process, nominated the encoding and decoding mechanism of genetic algorithm. Second, Ga process shall always contain a way to create a population, namely an initial set of potential solutions. Third, since Ga is based on the concept of natural selection, fitness selection from the environment is required for the system in order to evolve from the initial population. Hence, an evaluation function rating solutions in terms of their fitness value remains vital as well. Fourth, genetic operators that alter genetic composition of offspring plays an important part of controlling the evolving system such as crossover, mutation and selection. As the last part, parameter value setting including population size and probability of applying genetic operators is the last part to activate the genetic program.

In addition, based on some preference criterion/objective function, GA has been implemented to accelerate the speed of machine learning techniques from mixed media data widely as well (Mitra, Pal, & Mitra, 2002). For instance, researches have be done to validate of sensitivity to choice of parameters of the GA/KNN method based on Gene selection for sample classification based on gene expression data (Li, LP, 2001). The genes are able to be recognition subsequently to classify independent test set samples for the stochastic supervised pattern recognition model. In research of Support Vector Machines (SVM), Huang has proposed a GA based approach for feature selection enabling optimization of parameters (Huang & Wang, 2006). The results provide higher classification accuracy which has fewer input features for SVM. In deep learning studies, Momeni, Nazir, Armaghani, and Maizir (2014) has used artificial neural network (ANN) enhanced with genetic algorithm (GA) in finding global minima of predicting pile bearing capacity. Besides, Rouhi, Jafari, Kasaei, and Keshavarzian (2015) proposed a cellular neural network (CNN) whose parameters are determined by a genetic algorithm (GA) to solve benign and malignant breast tumors classification issues. Furthermore, Yang and Qin (2018) has developed a distributed correlation model mining from remote sensing big data based on gene expression programming (DCMM-GEP) in order to solve the enormous remote sensing data for mining algorithms. Based on genetic programming, the proposed DCMM-GEP have shown better R-square values, MAPE and less average time-consumption.

3. Methodology

The purpose of the paper is to construct an intelligent agent which consists of the prediction model for the required workforce with learning mechanism to solve workforce allocation issue for IC-D&S with precision and efficiency. This section includes three parts: (1) Problem Definition, which identifies the business scenario and the specific problems such as rule based constrains of the empirical study; (2) Data Preparation, which show that how the information flow and the related data be prepared in the first place for the further prediction and optimization model; (3) Model Construction for Intelligent Agent implies that how the learning based system been constructed. The implemented methodologies are described in detail in this section (Figure 1).

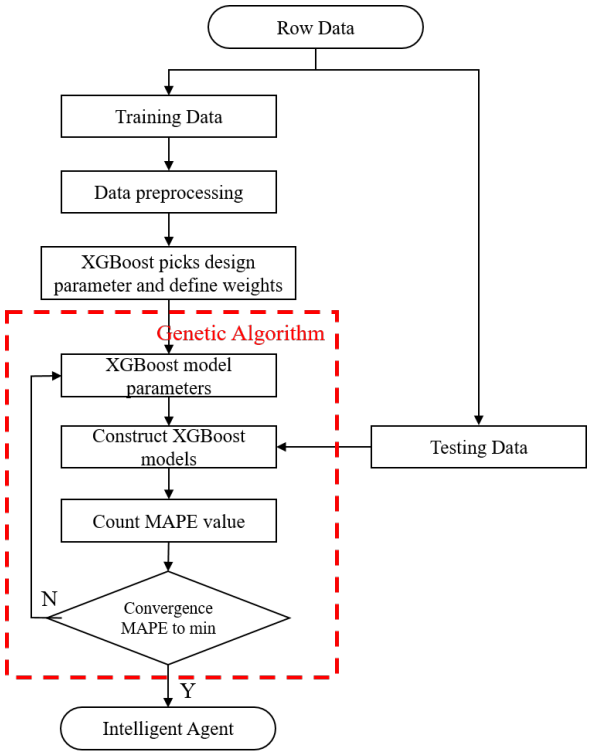

Figure 1

3.1 Problem Definition

Regarding as one of the most important resource for high-tech industries such as semiconductor industry, human capital management plays an important role to maintain enterprise's competitive advantages. Owing to the changing nature of knowledge workers in high-tech industries, jobs cannot be easily delineated. This means that conventional personnel selection methodologies which focus on static work and job analysis will no longer be appropriate for knowledge workers in high-tech industries (L.-F. Chen & Chien, 2011). Under the condition that routine work has gradually been replaced by machines, the expenditure of high-tech workers will become the critical part of the company's human capital costs. Yet seldom companies are capable to measure the productivity of workforce not to mention to quantify them. Considering the actual workforce required for each project, which usually depends on the historical experience of the company's internal high-level supervisors, and when the workforce estimates are inaccurate, it is likely that there will be problems of idle workforce or insufficient capacity, which indirectly leads to extra costs for the company. Besides, high cost of outsourcing is required to achieve a commitment to customer delivery. The purpose of this study is to solve the production deviation caused by the number of people making the project. The industry usually needs one department to spend one month to develop the manpower required for the whole year. When the number of projects is reduced, the study proposes a set. The decision support system effectively improves the efficiency and accuracy of manpower development. In this study, each the performance of workforce will be defined as the output of one manpower per day, nominated as Man-Day.

3.2 Data Acquisition

Data acquisition can be divided into two dependent parts including data preparation and data cleaning. In the data preparation part, it is necessary to collect the design parameters applied in the semiconductor industry to establish the correlation between design parameters and Man-Day via the discussion with field experts. Through the further information of the related chip design specifications, whether the factors have significant correlation with Man-Day, will be discussed in the subsequent model prediction. Finally, based on the company's statistical system, Man-Day data are required for each project of the project materials over the years which is collected and stored in the database of the company.

As the second part, data cleaning aims to filter corrupt or inaccurate records from a record set. General data cleaning process will filter outliers by boxplot, which divides continuous data into quartiles at a distance of 1.5 times the interquartile range of the first quartile or A point that is greater than the third quartile is divided into outliers. However, this method is less suitable, because there are many design parameters that are categorical variables. Hence, the study conducts the dominated method for data cleaning. Outliers are completely eliminated if they are dominated by other project. For example, if the design parameters of Project A are all more complicated than Project B, yet Project A presents Man-Day far less than Project B, and there exists Project C which is superior that both of Project A and Project B, the two materials will not be counted.

3.3 Model Construction for Intelligent Agent

The intelligent agent consists of two main parts including XGBoost and parameter optimization based on GA (Figure 1). Through the historical data obtained from the previous stage, the XGBoost machine learning model predicts important attribute factors that have a significant impact on the project and the relative impact of the factors on each design stage of the project. Detailed procedures are described as the follows. For the first part, collect all the project design parameters, and generate the sparse matrix by converting the project design parameters from category type to one hot encoding. Classification for the data into test set and training set are conducted so that it is available to establish DMatrix. DMatrix is an internal data structure used by XGBoost that is optimized for memory efficiency and training speed. XGBoost model was built using the training DMatrix obtained from pre-processing. XGBoost model has superior ability to import sparsity aware algorithms to process sparse matrices compared to many algorithms.

Table 1

General parameters	Booster parameters	Learning task parameters	Command line parameters
1. Booster	1. eta	1. objective	1. num_round
2. Silent	2. gamma	2. base_score	2. data
3. Verbosity	3. max_depth	3. eval_metric	3. test:data
4. Nthread	4. min_child_weight	4. seed	4. save_period
5. disable_defaul t_eval_metric	5. max_delta_step		5. task
	6. subsample		6. model_in
6. num_pbuffer	7. lambda		7. model_out
7. num_feature	8. alpha		8. model_dir
	9. tree_method		9. fmap
	10. sketch_eps		10. dump_format
	11. scale_pos_weight		11. name_dump
	12. updater		12. name_pred
	13. refresh_leaf		13. pred_margin
	14. process_type		
	15. grow_policy		
	16. max_leaves		
	17. max_bin		
	18. predictor		

In machine learning, parameters are the adjustable variables inside the model. The parameters in XGBoost model can be categorized into General parameters, Booster parameters, Learning task parameters and Command line parameters. Those detail parameters are as follow table. The common boosters are linear and tree respectively. It is proved by most studies that tree booster is better than linear booster. Therefore, in Booster parameters, this study will add

optimization algorithm to the booster of tree model via GA. GA adjusts the parameters to achieve the best predictive ability. The performance is evaluated by the fitness value in GA which is set to minimize the average absolute percentage error, and the parameter of the minimum average absolute percentage error is calculated as the benchmark parameter of the XGBoost machine learning model. In the agent system, there will be a field for the user to fill in the desired MAPE value. When the MAPE value is lower than the user-set benchmark, the gene algorithm will terminate by itself. When the user does not set any field, Then the gene algorithm will automatically iterate to defined sub-generations. If it cannot be detached, it is considered to be the best solution of the region.

4. Computation Results

In this section, the study implements the proposed intelligent agent system for an IC-D&S company in Taiwan as an empirical research case to solve the mentioned workforce allocation learning model. Parameter setting for XGBoost and GA for the model will be first described. Comparisons among the efficiency with the company's internal human decision making and the performance among different algorithms are conducted. For the sake of confidentiality, some conversion processing has been done, but it does not affect the conclusion and validity of the research.

The study applied the obtained row data from the company to train in the XGBoost machine learning model. The lifting method used in this study is tree booster. By the fix parameter setting for the model, the MAPE predicted by the following levels: (1) Stage 1 with MAPE = 27%, (2) stage 2 with MAPE = 28%, (3) stage 3 MAPE = 25% and (4) stage 4 with MAPE=22%. The study combines the results from the 4 stages to count total Man-Day of single project. The result shows that the MAPE for the project's total Man-Day is 26% in which we are able to determine this forecasting model is less accurate by the increasing scale of Man-Day's forecasting (Figure 2).

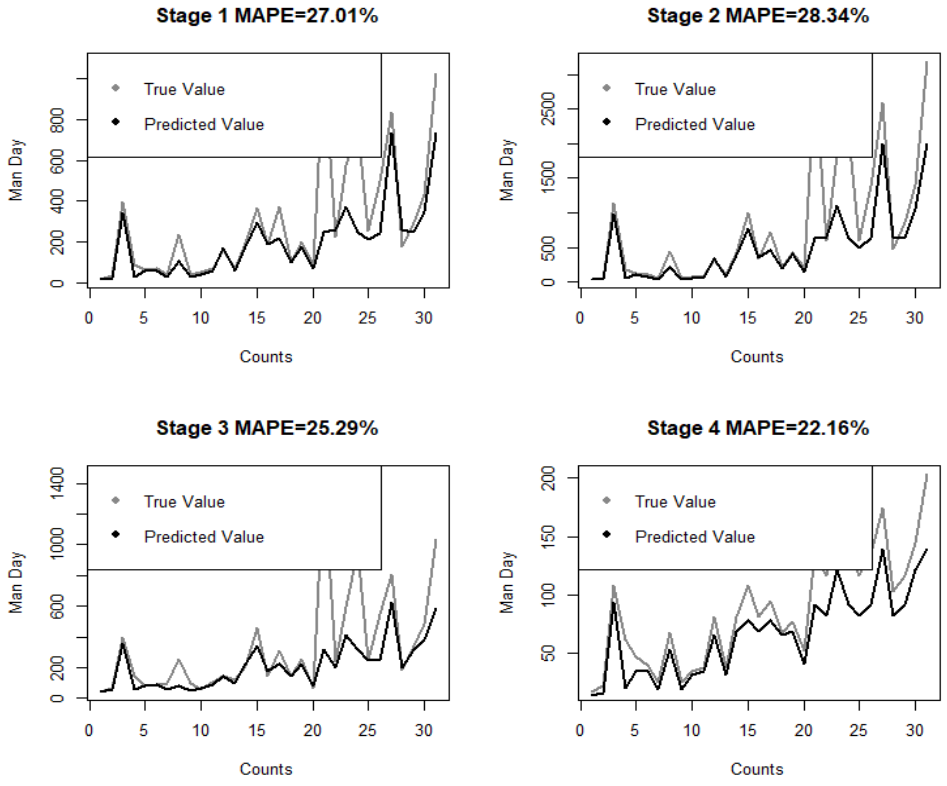

Figure 2

GA is used to optimize the Booster parameters in the XGBoost model. The real value type, fitness function for the booster parameters, lower bounds and upper bounds for booster parameters all follow the setting of XGBoost model. After setting crossover and mutation rate, the termination condition is set that if it does not evolve after 10 iterations,

which is viewed as the lowest MAPE and best predictive power for the XGBoost model parameters. The optimized XGBoost model parameters selects seven design parameters that have a significant impact on the project from 35 design parameters, and the relative impact of design parameters on each design stage of the project. Due to the protection of company secrets, the actual factor will not be shown, the table below (Table 2) shows which factor affects the time horizon forecast. The parameter-adjusted project predicted MAPE of 12.23%, which greatly increased the forecasting ability of the original unadjusted parameters.

Table 2

Design Parameter	Gain/Importance			
	Stage1	Stage2	Stage3	Stage4
Parameter A	0.9049	0.9493	0.9004	0.7870
Parameter B	0.0725	0.0317	0.0536	0.0892
Parameter C	0.0065	0.0088	0.0061	0.1176
Parameter D	0.0094	0.0054	0.0203	0.0038
Parameter E	0.0027	0.0022	0.0094	0.0012
Parameter F	0.0026	0.0018	0.0092	0.0006
Parameter G	0.0010	0.0004	0.0006	0.0004

This study also applies other algorithms in Boosting to conduct experiments. The three Boosting algorithms are GBM, XGBoost and LightGBM respectively. We initially compare the execution speed and prediction ability of the original model, in which GA are implemented as well. It is obvious that in this case, XGBoost's predictive ability is better than the original GBM model and the improved XGBoost version of lightGBM, but the program runtime is slightly higher than lightGBM, and far lower than GBM.

Table 3

	GBM	XGBoost	LightGBM
Original-MAPE	45.12%	26.36%	32.48%
Calculate execution time	4.0037sec	0.0509sec	0.0497sec

	GBM	XGBoost	LightGBM
Combine GA- MAPE	24%	12.23%	19.56%
Calculate execution time	189.3568sec	50.9552sec	46.2387sec

In this study, the important impact factors are determined by XGBoost machine learning model which is submitted to the domain experts for professional knowledge evaluation. It is found that the seven attributes selected by the intelligent agent plays an important part for workforce allocation in IC service design project, and the case company provides the existing The historical data of the project forecast, the case company's ability to predict by manual is MAPE=39%, and the prediction model proposed in this study can improve the prediction accuracy for at least 3 times.

To sum up, based on the proposed intelligent agent for workforce allocation and consumption forecasting, the system can determine the new training data set and change the XGBoost model with GA to ensure the accuracy of the system prediction and response to the case company technology and services in time. In order to ensure that the domain knowhow of decision makers or experts can also be considered in the machine learning model, this study additionally adds a user input system that changes the original data set whenever the decision makers are likely to change the person rate manually. For cases that the estimated value is far from the prediction of the professional domain knowledge, the decision maker can change the manpower to use the predicted value and input the data into the original data set to establish a new machine learning model. This decision support system can give decision makers the basis for pre-planning manpower and time schedule. The MAPE of the forecasting result is 13.39%. After adding new information to the intelligent agent, the MAPE will be reduced significantly which is better for decision makers for productivity management.

5. Conclusion

IC Design has been an industry which provides flexible application-specific integrated circuit (ASIC) services enabling semiconductor manufacturing companies for flexible decision. The decision-making process for standardizing manpower planning will greatly help the industry. It will ensure that the company can produce projects and research and development technology with the most appropriate manpower, thereby maintaining the company's high degree of competitiveness in terms of financial or technical aspects. For IC design service industry, the main productivity denotes to IC design which is influenced by the performance of project management from workforce allocation. This study proposes an intelligent agent system for standardizing IC-D&S project process, which combines machine learning algorithms and analyzing from the existing data. The study trains a XG Boosting model with Genetic Algorithm (GA) based parameter optimization mechanism combining the existing rules of thumb of domain experts to optimize by time and reduce the bias of human estimation. The proposed intelligent agent contributes to Total Resource Management (TRM) to enhance productivity, reduce costs and intelligence management.

Acknowledgements

This research is supported by Ministry of Science and Technology, Taiwan (MOST 107-2634-F-007-002; MOST 107-2634-F-007-009).

References

Booker, L. B., Goldberg, D. E., & Holland, J. H. (1989). Classifier systems and genetic algorithms.

Box, G. E., & Hunter, J. S. (1957). Multi-factor experimental designs for exploring response surfaces. *The Annals of Mathematical Statistics, 28*(1), 195-241.

Chen, L.-F., & Chien, C.-F. (2011). Manufacturing intelligence for class prediction and rule generation to support human capital decisions for high-tech industries. *Flexible Services and Manufacturing Journal, 23*(3), 263-289.

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree boosting system.* Paper presented at the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

Chen, Y.-J., & Chien, C.-F. (2018). An empirical study of demand forecasting of non-volatile memory for smart production of semiconductor manufacturing. *International Journal of Production Research*, 1-15.

Chien, C.-F., Chen, W.-C., & Hsu, S.-C. (2010). Requirement estimation for indirect workforce allocation in semiconductor manufacturing. *International Journal of Production Research, 48*(23), 6959-6976.

Chien, C.-F., Chen, Y.-J., Hsu, C.-Y., & Wang, H.-K. (2014). Overlay error compensation using advanced process control with dynamically adjusted proportional-integral R2R controller. *IEEE Transactions on Automation Science and Engineering, 11*(2), 473-484.

Chien, C.-F., Chen, Y.-J., & Peng, J.-T. (2010). Manufacturing intelligence for semiconductor demand forecast based on technology diffusion and product life cycle. *International Journal of Production Economics, 128*(2), 496-509.

Chien, C.-F., Dauzère-Pérès, S., Ehm, H., Fowler, J. W., Jiang, Z., Krishnaswamy, S., . . . Uzsoy, R. (2011). Modelling and analysis of semiconductor manufacturing in a shrinking world: challenges and successes. *European Journal of Industrial Engineering 4, 5*(3), 254-271.

Chien, C.-F., Hong, T.-y., & Guo, H.-Z. (2017). A Conceptual Framework for "Industry 3.5" to Empower Intelligent Manufacturing and Case Studies. *Procedia Manufacturing, 11*, 2009-2017.

Darwin, C. (2004). *On the origin of species, 1859*: Routledge.

Friedberg, R. M. (1958). A learning machine: Part I. *IBM Journal of Research and Development, 2*(1), 2-13.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1): Springer series in statistics New York, NY, USA:.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367-378.

Holland, J. H. (1975). Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence: University of Michigan Press Ann Arbor.

Huang, C.-L., & Wang, C.-J. (2006). A GA-based feature selection and parameters optimizationfor support vector machines. *Expert Systems with applications, 31*(2), 231-240.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T.-Y. (2017). *Lightgbm: A highly efficient gradient boosting decision tree.* Paper presented at the Advances in Neural Information Processing Systems.

Kearns, M. (1988). Thoughts on hypothesis boosting. *Unpublished manuscript, 45*, 105.

Khakifirooz, M., Chien, C. F., & Chen, Y.-J. (2018). Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0. *Applied Soft Computing, 68*, 990-999.

Koza, J. R. (1994). Genetic programming as a means for programming computers by natural selection. *Statistics and computing, 4*(2), 87-112.

Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., & Hoffmann, M. (2014). Industry 4.0. *Business & Information Systems Engineering, 6*(4), 239-242.

Leachman, R. C., Ding, S., & Chien, C.-F. (2007). Economic efficiency analysis of wafer fabrication. *IEEE Transactions on Automation Science and Engineering, 4*(4), 501-512.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997a). The bullwhip effect in supply chains. *Sloan management review, 38*, 93-102.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997b). Information distortion in a supply chain: The bullwhip effect. *Management science, 43*(4), 546-558.

Lee, J., Bagheri, B., & Kao, H.-A. (2015). A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing Letters, 3*, 18-23.

Lee, J., Kao, H.-A., & Yang, S. (2014). Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia Cirp, 16*, 3-8.

Michalewicz, Z. (1996). Evolution strategies and other methods *Genetic Algorithms+ Data Structures= Evolution Programs* (pp. 159-177): Springer.

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: a survey. *IEEE transactions on neural networks, 13*(1), 3-14.

Momeni, E., Nazir, R., Armaghani, D. J., & Maizir, H. (2014). Prediction of pile bearing capacity using a hybrid genetic algorithm-based ANN. *Measurement, 57*, 122-131.

Nielsen, D. (2016). *Tree Boosting With XGBoost-Why Does XGBoost Win" Every" Machine Learning Competition?* , NTNU.

Rouhi, R., Jafari, M., Kasaei, S., & Keshavarzian, P. (2015). Benign and malignant breast tumors classification based on region growing and CNN segmentation. *Expert Systems with applications, 42*(3), 990-1002.

Schaller, R. R. (1997). Moore's law: past, present and future. *IEEE spectrum, 34*(6), 52-59.

Schapire, R. E., & Freund, Y. (2012). *Boosting: Foundations and algorithms*: MIT press.

Wu, J.-Z., & Chien, C.-F. (2008). Modeling semiconductor testing job scheduling and dynamic testing machine configuration. *Expert Systems with Applications, 35*(1-2), 485-496.

Yang, L., & Qin, Z. (2018). Distributed correlation model mining from remote sensing big data based on gene expression programming. *Peer-to-Peer Networking and Applications, 11*(5), 1000-1011.

Biographies

Chieh Hsu is a master student in Industrial Engineering and Engineering Management in National Tsing Hua University, Hsinchu, Taiwan. She had the related experience in IC Design Service and semiconductor industry. As a member of Artificial Intelligence for Intelligent Manufacturing Systems Research Center, MOST, Taiwan, Chieh Hsu's main research field is about applying artificial intelligence and big data analysis in empirical studies.

Hsuan-An Kuo is a PhD student of Industrial Engineering and Engineering Management in National Tsing Hua University, Hsinchu, Taiwan. Hsuan An had the experience in both traditional and high technology industry. For recent research, he dedicates in implementing system simulation and optimization methodology in semiconductor supply chain issues.

Ju-Chien Chien is a senior student in Computer Science Department of National Tsing Hua University, Hsinchu, Taiwan. She has been an intern in Global Unichip and Artificial Intelligence for Intelligent Manufacturing Systems Research Center, MOST, Taiwan.

Wenhan Fu is PhD candidate in National Tsing Hua university, Taiwan. His research interests include demand forecast, data analytics, supply chain management and smart production.

Kang-Ting Ma is a postdoctoral researcher with the Artificial Intelligence for Intelligent Manufacturing Systems (AIMS) Research Center in National Tsing Hua University, Hsinchu, Taiwan. He received his Ph.D. in Industrial Engineering and Engineering Management from National Tsing Hua University (NTHU), Hsinchu, Taiwan. His works have been published in European Journal of Operations Research. His research interests include operations research, decision analysis, and smart production.

Chen-Fu Chien is currently a Tsinghua Chair Professor and a Micron Chair Professor with NTHU. He is also the Director of the Artificial Intelligence for Intelligent Manufacturing Systems Research Center sponsored by the Ministry of Science and Technology, the NTHU-TSMC Center for Manufacturing Excellence, and the Principal Investigator for the Semiconductor Technologies Empowerment Partners (STEP) Consortium. He holds eight U.S. invention patents on semiconductor manufacturing. His research mainly concerns the development of better analytical methods including big data analytics, decision analysis, and optimization algorithms and solutions for high-tech companies confronting with decision problems involved in strategy, manufacturing, and technology that are characterized by uncertainty and a need for tradeoff among various objectives and justification for the decisions. His publication number is up to 170, and his publication has been cited for 5057 times. He has a number of case studies in Harvard Business School. He proposed Industry 3.5 as a hybrid strategy between the existing Industry 3.0 and to-be Industry 4.0, empowered by AI and big data analytics for disruptive innovations. His book on Industry 3.5 (ISBN 978-986-398-380-4) is a bestselling book.