2018

Ensemble Machine Learning Systems for the Estimation of Steel Quality Control

Fucun Li
University of Wollongong, Jiangsu Jinheng Information Technology Co., Ltd, fl626@uowmail.edu.au

Jianqing Wu
University of Wollongong, jw937@uowmail.edu.au

Fang Dong
Southeast University, fdong@seu.edu.cn

Jiayin Lin
University of Wollongong, jl461@uowmail.edu.au

Geng Sun
University of Wollongong, gsun@uow.edu.au

See next page for additional authors

Ensemble Machine Learning Systems for the Estimation of Steel Quality Control

Abstract

Recent advances in the steel industry have encountered challenges in soliciting decision making solutions for quality control of products based on data mining techniques. In this paper, we present a steel quality control prediction system encompassing with real-world data as well as comprehensive data analysis results. The core process is cautiously designed as a regression problem, which is then best handled by grouping various learning algorithms with their massive resource of historical production datasets. The characteristics of the currently most popular learning models used in regression problem analysis are as well investigated and compared. The performance indicates our steel quality control prediction system based on ensemble machine learning model can offer promising result whilst delivering high usability for local manufacturers to address the production problem by aid of development of machine learning techniques. Furthermore, real-world deployment of this system is demonstrated and discussed. Finally, future directions and the performance expectation are pointed out.

Disciplines

Engineering | Science and Technology Studies

Publication Details

Li, F., Wu, J., Dong, F., Lin, J., Sun, G., Chen, H. & Shen, J. (2018). Ensemble Machine Learning Systems for the Estimation of Steel Quality Control. 2018 IEEE International Conference on Big Data (Big Data) (pp. 2245-2252). United States: IEEE.

Authors

Fucun Li, Jianqing Wu, Fang Dong, Jiayin Lin, Geng Sun, Huaming Chen, and Jun Shen

Ensemble Machine Learning Systems for the Estimation of Steel Quality Control

Fucun Li[1, 3], Jianqing Wu[1], Fang Dong[2], Jiayin Lin[1], Geng Sun[1], Huaming Chen[1], Jun Shen[1]
[1] *School of Computing and Information Technology*
University of Wollongong, Wollongong, Australia
[2] *School of Computer Science and Engineering*
Southeast University, Nanjing, Jiangsu, P.R. China
[3] *Jiangsu Jinheng Information Technology CO., LTD, Nanjing, Jiangsu, P.R. China*
Email: {fl626, jw937, jl461, hc007}@uowmail.edu.au, {gsun, jshen}@uow.edu.au, fdong@seu.edu.cn

***Abstract*—Recent advances in the steel industry have encountered challenges in soliciting decision making solutions for quality control of products based on data mining techniques. In this paper, we present a steel quality control prediction system encompassing with real-world data as well as comprehensive data analysis results. The core process is cautiously designed as a regression problem, which is then best handled by grouping various learning algorithms with their massive resource of historical production datasets. The characteristics of the currently most popular learning models used in regression problem analysis are as well investigated and compared. The performance indicates our steel quality control prediction system based on ensemble machine learning model can offer promising result whilst delivering high usability for local manufacturers to address the production problem by aid of development of machine learning techniques. Furthermore, real-world deployment of this system is demonstrated and discussed. Finally, future directions and the performance expectation are pointed out.**

Keywords— ensemble learning, steel quality control, intelligent manufacturing, data mining

I. Introduction

Steel is an essential raw material for industrial manufacturing. It is widely used everywhere, including buildings, bridges, ships, containers, medical instruments, and automobiles. In the steel industry, quality control of products is critical, which is determined by the characteristics of the industry. As the steel industry is a process-oriented industry, each production line is continuously produced in the production process. Furthermore, the production scale of each production line is relatively large, which means, a whole batch of products would be impacted with similar quality problems while quality problems occur. Under this circumstance, it will result in severe economy losses. Thus, companies in the steel industry seeks to solve this risk by introducing a better quality control solution, which is recently termed as intelligent manufacturing.

The steel production process also is very complicated. The entire procedure mainly consists of five stages of processes, which are iron-making, steel-making, hot-rolling, cold-rolling and heat-treatment [1]. The quality control of steel production is also very complicated during each stage of the process. It is a major challenge for the entire steel industry. The conventional approaches mainly use laboratory equipment to verify quality data. It refers to sampling the product in different production processes. The samples are returned to the laboratory and processed by the laboratory instruments for analysis. For example, a spectrum analyser is used to determine the chemical composition of the molten steel while a tensile machine is applied to determine the tensile strength of the finished product. Once obtaining the test results, the production will be carried out in terms of the plan if the results meet corresponding to the users' requirements. If it is not satisfied, it will be required to be re-processed in a subsequent process. Thus, these conventional test methods are not only cost-intensive but also extremely time-consuming. These impact the effectiveness and efficiency of the manufacture of factory.

An alternative option is to deploy statistical processing control (SPC) system, which has become widely used in the iron and steel industry [2]. However, the SPC system can only give warning of the parameters in the production process, and it cannot predict the actual values relevant to product quality [3]. Currently, the integration of IoT, big data, cloud computing and artificial intelligence technologies help manufacturing industries implement manufacturing cyber-physical systems, which is one of the critical features of Industry 4.0 [4]. Additionally, the development of machine learning, including various algorithms and theories, allows the researchers and developers to deal with the demands of manufacturing data analysis [5]. Machine learning methods have great potential to discover knowledge out of the vast amounts of data with the sustainable increase of manufacturing data repositories [4] [6] [7].

Regarding the steel quality control problem, there is by far not elegant solution from both system level design and computational model deployment. One major issue is considered as the lack of on-site real data issue and the other is the domain knowledge from a long-term engagement in the steel industry. In this paper, we target this steel quality control predicting problem from both to present a comprehensive solution. The goal of predicting outcomes and uncovering the latent relationships in data is to turn massive amounts of *manufacturing data* into valuable information and knowledge that support the manufacturing system to improve decision-making process [8] [9] [10]. Lastly, a real-world deployment in

Nanjing Iron and Steel company is successfully implemented at the early stage of the steel production process.

The rest of this paper is organized as follows: the background and related work will be studied in section II; following the ensemble method for steel quality control will be discussed in section III, where the system framework for steel quality control will be also provided; an overall evaluation of the real-world deployment in the Nanjing Iron and Steel company is presented in section IV while the model performance is collected and discussed in section V; section VI concludes the paper.

II. Background

In this section, we briefly review how the steel production process runs in the steel industry, and revisit the statistical process control for the estimation of steel quality control. Finally, we introduce the ensemble machine learning.

A. Steel production process

The industrial steel production process is demonstrated in Figure 1, consisting of ironmaking, steelmaking, hot rolling, cold rolling and heat-treatment. Briefly, the first process is to obtain raw materials such as iron ores and coking coals, and then refine them into the cast irons in the ironmaking blast furnace. Secondly, the cast irons are smelted into steels using steel-making furnaces. Thirdly, the steels should be cast into steel ingots or slabs, and then be transported to the rolling mill for rolling or forging. Finally, steels can be molded to various shapes.

The steel production requires a series of manufacturing procedures. Given each procedure is complicated, time-consuming and expensive, each of the links needs to enforce strict quality control. The information systems of steel companies generate and accumulate a significant amount of data, which can be considered as input variables [11]. Moreover, the input variables collected from the steelmaking processes are quite noisy [12]. The data cleaning and prepossessing should be undertaken to deal with inaccurate data and the specific use. Additionally, to describe the target concept sufficiently, feature selection is necessary in the system to reduce the dimensions of the datasets and increase model accuracy[13].

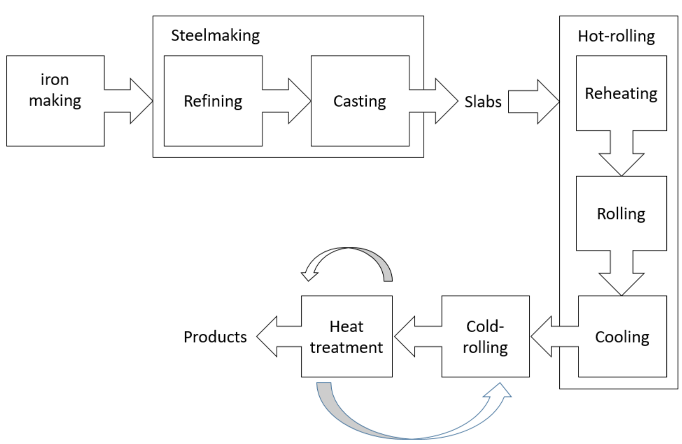

Figure 1. Schematic diagram of Steel Production processes [1]

B. Statistical Process Control

SPC is a quality control method, based on cause-effect relationships, for monitoring and controlling the quality of the manufacturing processes through the reduction of variability[14] [15] [16]. It plays a vital role in improving the competitiveness of their products in the steel industry [17]. The conventional monitoring method, Univariate Statistical Process Control (USPC) cannot detect abnormality easily when we have a large number of observations [1]. Multivariate Statistical Process Control (MSPC) can handle the high-dimensional and correlated variables in the process through reducing the dimension of process variables and decomposing the correlation between them; principal component analysis (PCA) and the partial least squares (PLS) both are the most widely used in MSPC methods [18].

C. Problem Description

The accurate prediction of the physical properties of steel has become an important research problem in the steel industry. Due to some limitations of conventional USPC and MSPC, Data mining techniques offer practical solutions for conventional USPC and MSPC methods limitation, such as the emphasis on diagnosis rather than detection, preprocess massive and multiple source datasets, complicated process and non-parametric problems [1] [12] [19] [20].

Regression analysis is utilized to estimate the relationships between a combination of input variables (dependent) and an outcome variable (independent); it functions in understanding how the typical value of the dependent variable (actual value) changes when an independent variable (predicted value) varies [21].

The main goal is to explore the prediction of performance in the steel production process by using continuous variables, such as temperature and pressure information. These variables are regularly recorded in the system. The prediction system is achieved by defining the task as a traditional regression problem, which involves applying one or more continuous inputs to forecast a desired output [22]. In this formulation, the estimation of steel quality control can be regarded as an objective output of the regression analysis.

III. Ensemble Machine Learning for Steel Quality Control

A. Ensemble Methods

Ensemble methods are learning algorithms that combine multiple differential machine learning models to improve the prediction performance [23] [24]. A set of machine learning models can be referred to as base learners or weak learners [25]. Generally, an ensemble method helps to achieve stronger generalisation ability than that of a single model [25] [26]. Bagging, boosting, stacked generalisation (stacking) are the three most common approaches used to generate ensemble systems for solving different problems [23]. Stacking is a

specific combination method, which combines a set of the first-level individual learners as input to a meta-learner for enhancing prediction accuracy and robustness [27] [28].

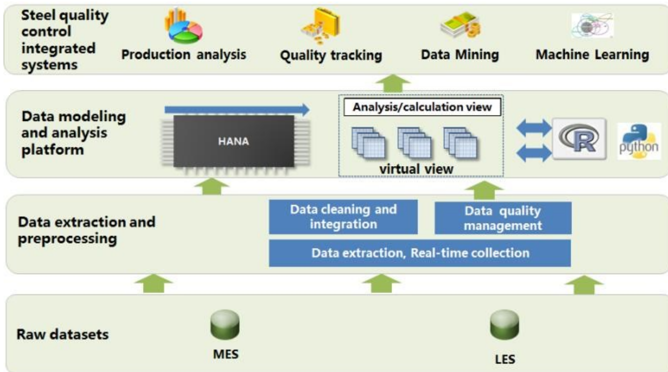

Figure 2. System Framework

B. *System Framework for Nanjing Iron and Steel Company*

The ensemble algorithm as well as the associated data flow has been implemented in a comprehensive system of Nanjing Iron and Steel Company. Figure 2 demonstrates the framework of this machine learning-based performance prediction system, which consists of four major layers from bottom to the top: raw datasets, data extraction and preprocessing, data modelling and analysis platform, and steel quality control integrated systems.

Raw datasets contain historical observations, a manufacturing execution system (MES) and Lab Execution System (LES) information.

Data extraction and preprocessing include analysis and processing of the abnormal data, filling of the missing values, and duplicate data removal. Due to the difference between the feature variables, to accelerate the convergence of the model, the processing of the datasets should be scaled their feature variables down to a range between 0 and 1.

Data modelling and analysis platform use Extract Transform Load (ETL) tools and real-time acquisition function to extract the information from the historical data set, which is imported into the system applications and products high-performance analytic appliance (SAP HANA) based data warehouse. The HANA built-in analysis tools, R language, and Python are also included to establish the platform.

Finally, based on the data modelling and analysis platform, production process analysis, quality tracking, data mining, and machine learning are developed as several major components to form the complete steel quality control layer.

C. *Working Principle and Data Flow in System*

As the steel quality control system using ensemble learning has been developed and implemented to solve the regression problem, Figure 3 shows a flowchart of the working principle of the ensemble learning system. After obtaining the MES and LES data from the steel quality management system, data analysis, data processing, and feature engineering are conducted to select 57 features as an input to the estimation

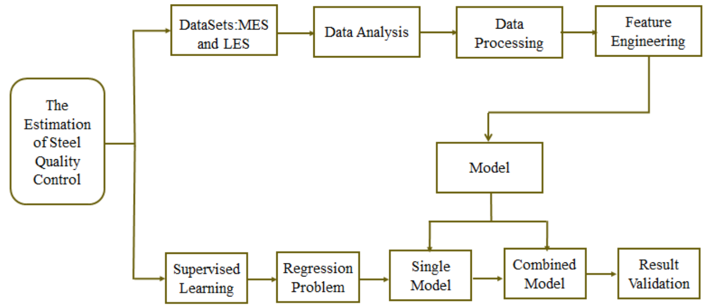

Figure 3. Flow Chart of Ensemble Learning System

models. As can be seen from Figure 4, we also apply a heatmap to rank the top 10 correlation features. A high values demonstrate a strong relationships between features.

Provided that the data extraction and preprocessing have been well conducted, sorted data are then fed into the layer above it. Specifically, the data analysis phase is to understand the relationship between raw data and to explore their correlations and distribution. At the data processing phase, removing all duplicate records and replacing missing data process with valuable data are executed accordingly, which can be then utilized to provide accurate predictions for the steel company. Finally, based on a series of domain knowledge, we select 57 features such as the thickness of steel plates, the rolling temperature, the amount of cooling water, and chemical elements.

Moreover, as the estimation of steel quality control is a supervised regression learning, we start with the core machine learning concepts to generate a set of T base learners $\{h_1, \ldots h_T\}$ to tackle the problem. In other words, the original training data is divided in a manner of T-fold. Then, the

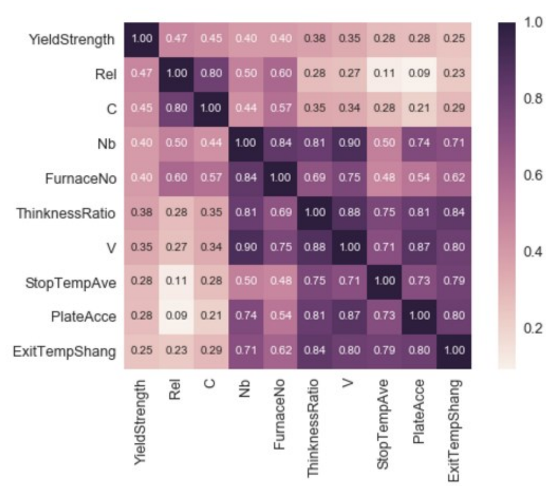

Figure 4. Rank Top 10 Features in the Correlation Heatmap

training data is processed for T times and each time it only produces one prediction.

In the top layer of the system, averaging method and stacking method works together with the initial learners in order to form a combined model, which functions in obtaining the corresponding combined output $H(x)$ for the dependent variable x respectively. Furthermore, for the result validation stage, we measure our proposed model combinations by R-squared (R2), Root Mean Square Error (RMSE), and Percentage of Error (PE).

D. Data Availability

Most steel companies have established information systems at the earlier stage, such as MES, LES, Enterprise Resource Planning (ERP), Energy Management System (EMS), and supervisory control and data acquisition (SCADA). The systems house the data generated during the whole product lifecycle. It includes order data, quality information in product manufacturing processes, price and quality information of the raw materials, process parameter information, as well as the energy consumption information from the production process. The datasets describe the entire life cycle of corresponding products. In this study, the data sources are primarily collected from the existing information systems of the Nanjing Iron and Steel company, which are MES and LES.

IV. EMPIRICAL EVALUATION

A. Datasets.

We use two different sets of data as demonstrated in Table 1. The detailed explanation is as follows.

- **MES dataset:** It is an information management system for the production process execution layer of the manufacturing enterprise. We mainly obtained production process data of steel plates and process data from the MES system of the Nanjing Iron and Steel Company, including rolling performance, cooling performance, continuous casting performance and chemical composition.
- **LES dataset:** LES refers to the inspection and testing system of a manufacturing enterprise. We mainly obtained the performance evaluation data of the steel plate from the LES system, including yield strength, tensile strength, elongation and impact work.

B. Baselines.

In this paper, we extend the comparison of our deployed ensemble models with the following baselines:

- LM: Linear models learn functions predicted by a linear combination of attributes. Linear models include Linear Regression, Ridge Regression, Lasso Regression and

TABLE I. DATASETS

Dataset	MES	LES
Data source	Enterprise production process execution system	Inspection and testing system
Time Span	2014/4-2017/8	2014/4-2017/8
Granular Data	Rolling performance parameter, cooling performance parameter, continuous casting performance parameter, and chemical composition parameter.	Sample serial number
Data detailed information	The rolling performance of steel plates (Obtained by the Steel plate number) includes 17 characteristics such as rolling mode, number of passes, rolling temperature, final rolling temperature, and thickness of the intermediate blank.	The Inspection and testing data include four indexes of yield strength, tensile strength, elongation and impact energy of the steel plate.
	The cooling performance of steel plates (Obtained by the Steel plate number) includes 13 characteristics such as water temperature, water pressure, water inlet temperature, water volume and reddening temperature.	
	The continuous casting performance and chemical composition (Obtained by the slab number) include 25 essential characteristics, such as chemical composition: C, Mn, P, S, Si and other main chemical components. The continuous casting performance: medium temperature, drawing speed and average liquid level.	

TABLE II. COMPARISONS WITH BASELINES ON MES AND LES

Model	R^2	RMSE	Percentage of error (%)
Linear Regression	0.431	17.774	3.22
Ridge Regression	0.443	17.576	3.18
Lasso Regression	0.444	17.580	3.18
Elastic Net	0.443	17.588	3.18
SVM(RBF)	0.522	16.288	2.95
KRR(RBF)	0.539	15.993	2.90
KNN (distance)	0.530	16.158	2.93
RF	0.530	16.151	2.92
GBDT	0.539	15.996	2.90
LGBM	0.540	15.973	2.89
XGBoost	0.545	15.890	2.88

ElasticNet.

- SVM(RBF): Support Vector Machine is a widely used machine learning model for both classification and regression problems. The selection of kernel function is a critical factor concerning the performance, among which Radial Basis Function (RBF) kernel function is widely used [29].
- KRR: Kernel Ridge Regression is a well-known model combining Ridge Regression with the kernel function.
- KNN: K-Nearest Neighbours is non-parametric model incorporating with the distance matrix between data samples. KNN is flexible in distance matrix definitions and mostly its performance is good in practice.
- EL: Ensemble Learning is a powerful machine learning paradigm to use the multiple models for decreasing variance (bagging), bias (boosting) or making accurate predictions (stacking) [25].
- RF: Random Forest is one of the most popular bagging algorithms, which is based on the decision tree algorithm [23]. RF has better resistibility to overfitting and usually has less variance.
- GBDT: Gradient Boosting Decision Tree is an ensemble model, which train a series of decision trees sequentially [30] [31].
- LGBM: Light GBM is a variant of Gradient Boosting Machine (GBM), which is a novel GBDT algorithm with Gradient-based One-Site Sampling (GOSS) and Exclusive Feature Bundling (EFB) techniques to deal with data instances and features respectively [30] [32].
- XGBoost: eXtreme Gradient Boosting is an efficient and scalable implementation for tree boosting [33] [34].

V. MODEL PERFORMANCE

A. Evaluation Measurements

To evaluate the performance of our prediction system and furthermore to compare with our baselines, R-Square and root mean square error are included as the main measurements.

In details, we firstly define that y_i is the observed value, $\bar{y}$ is the average of the observed values, $\hat{y}_i$ is the predicted value, and n is the number of all available ground truths. $\bar{y}$ is derived from Equation (1), relying on the availability of the observed value y_i :

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1}$$

Several evaluation measurements are defined as below, Equation (2-4):

R-Square: the improvement in prediction from the regression model compared to the mean model. This value ranges from 0 to 1, while a closer value to 1 indicates a better prediction results.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

Root Mean Square Error (RMSE): the standard deviation of the residuals, measures the differences between the values which predicted by a model and the values observed. The lower the RMSE value, the better prediction results we get.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$\text{PE} = \frac{RMSE}{\bar{y}} \tag{4}$$

Another metrics, percentage of error (PE), is defined in Equation (4) to illustrate the percentage of error comparing with the average of the observed values. As shown in Equation (4), it stays synchronous with the RMSE level.

The experiment results in the execution of our 11 baselines are shown in Table 2. Without any doubts, we can find that algorithms involve boosting strategy (GBDT, LGBM, and XGBoost) perform better than conventional single regression models (Linear Regression, Ridge Regression, and Lasso Regression). This result further indicates that the ensemble model can have the potential to outperform a single regression model.

The implementation of the ensemble learning system is carried out in the following sequence: training the prediction of a set of single models (base learners) at the *first-level learner* stage, the prediction of combined model (meta-learner) at *second-level learner* stage, and then the performance evaluation [27].

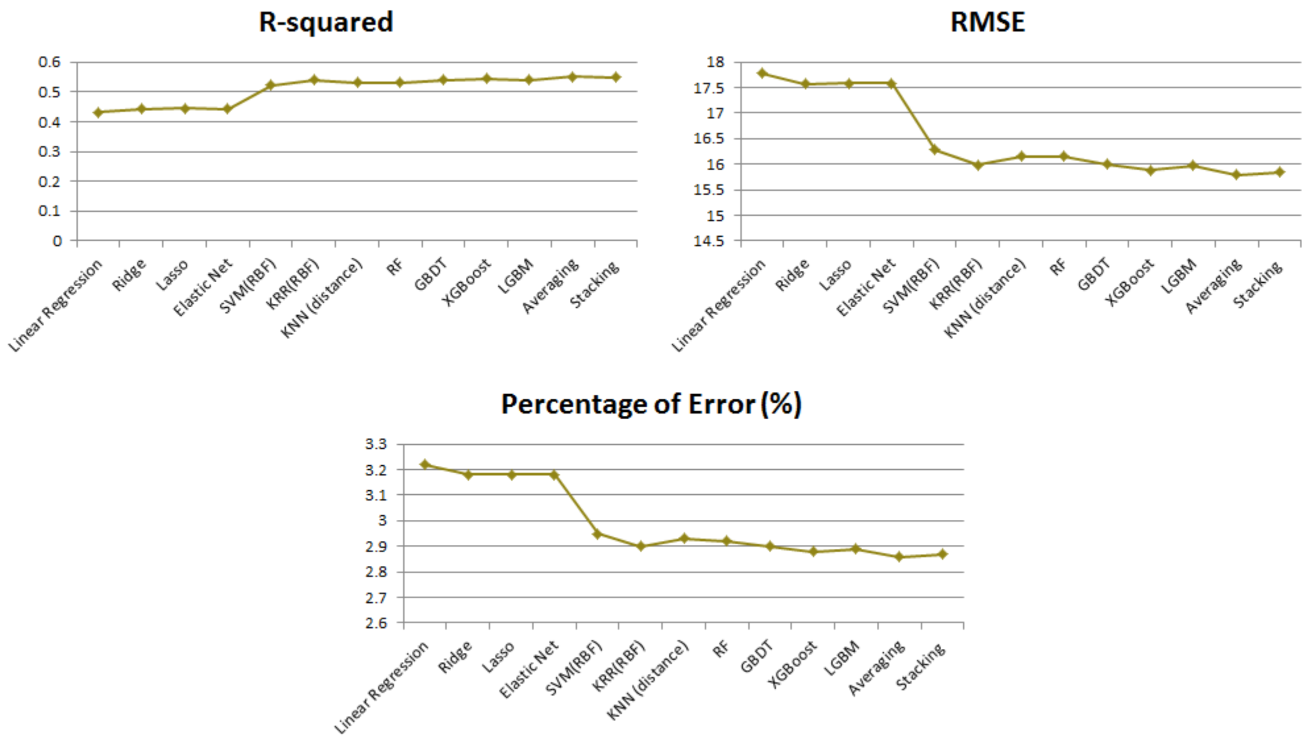

Figure 5. Model Results on MES and LES.

B. Main findings

To improve the accuracy of the model and reduce the over-fitting, we adopt the methods of averaging and stacking sequentially. Based on the “many could be better than all” theorem, we only include some single learners to compose an ensemble model instead of applying all of them, which obtains superior performance [25] [35].

Additionally, to reach a desired ensemble model, the selected base learners should be as diverse as possible between each other, while they can also report good performance independently [25] [36]. In this work, the combination strategy is as follows:

1) Combine the models with the Averaging method and combine the following models: SVM, KRR GBDT, XGBOOST.

2) Combine the models with the Stacking method and combine the following models SVM, KRR, lightGBM, XGBOOST.

The results of the Averaging ensemble model and Stacking ensemble model are illustrated in Table 3 and Figure 5. We reported all results regarding the base models as well as the ensemble models in Figure 5 and Figure 6.

From Table 3 and Figure 5, we can see that the two assembled models outperform the single ones from all the evaluation metrics, including R^2, RMSE and percentage of error (PE). A lower RMSE and a lower PE indicate a better model whilst vice versa for R^2.

From the statistics and plots, it is easy to see that the assembled models, averaging ensemble model and stacking ensemble model, have lower RMSE and PE values and higher R-square values.

TABLE III. COMPARISONS WITH AVERAGING MODEL AND STACKING MODEL

Model	**Averaging Model**	**Stacking Model**
R^2	0.550	0.548
RMSE	15.790	15.847
Percentage of error (%)	2.86	2.87

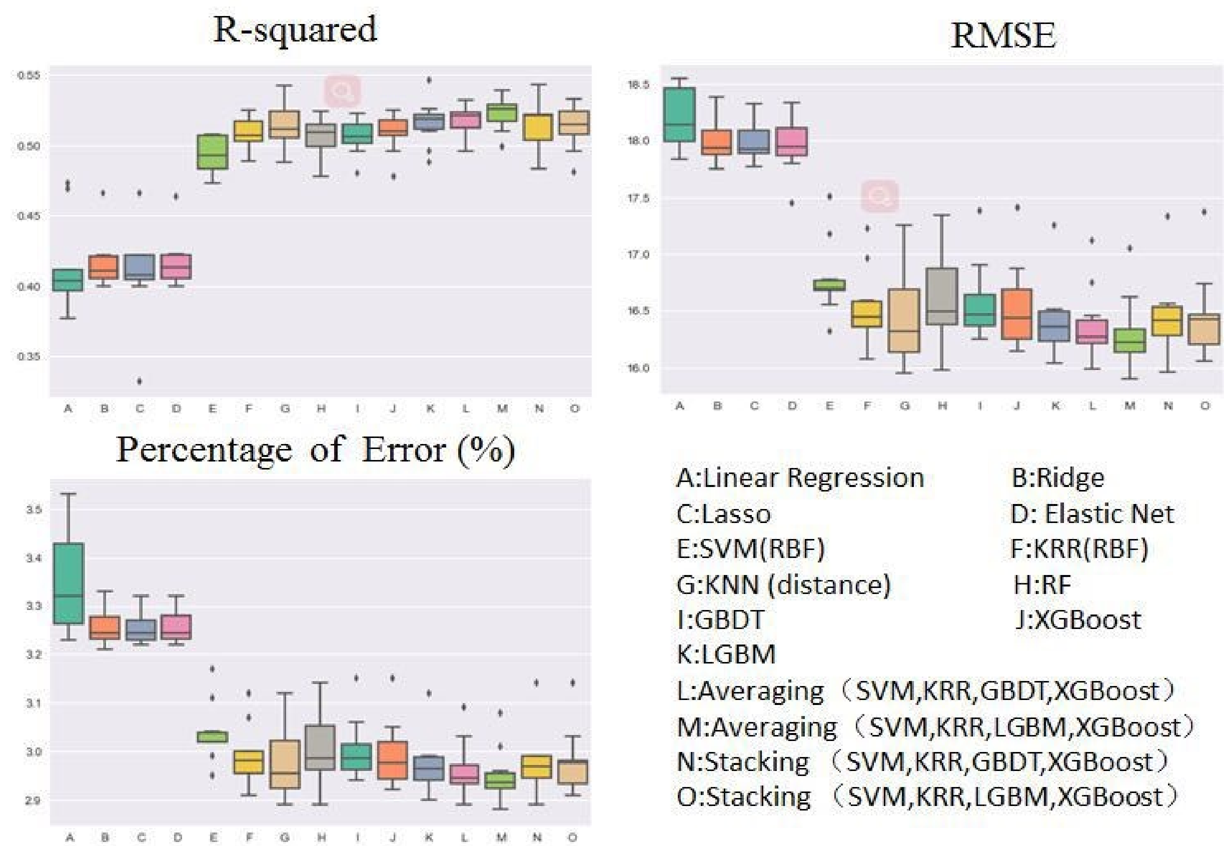

Figure 6. Models Ten Folds Random Evaluation

To better measure the performance of the models, the prediction errors of the models are collected in ten folds random evaluation. ‘L’, ‘M’, ‘N’ and ‘O’ is reported as the averaging ensemble and stacking ensemble models in Figure 6. As we can see from Figure 6, the averaging combination of SVM, KRR, LGBM, XGBoost (M) is better than other stacking method and single models.

By far, 4905 real-world samples from the on-site collection is used in our models for predicting steel quality. It is worth to consider that the model accuracy will be further improved when a larger and more comprehensive dataset is continuously generated and fed into our models in the near future.

VI. Conclusion ANd Future Work

In this paper, we propose a practical machine learning enabled prediction system for forecasting steel quality control, based on historical observation data. We evaluate the system, specifically concerning the performance of the deployed ensemble models on two types of datasets: MES and LES. The ensemble machine learning systems for steel quality control achieves performances which are significantly higher than other 11 baseline approaches, confirming that our prediction system is better and more robust to the steel quality prediction.

So far, this successfully deployed steel quality prediction system lays a foundation to incorporate machine learning and data analytics technologies in the real-world production process. The design of this system has taken the advantages from the data collection, domain knowledge building as well as the machine learning technologies. Therefore, there is great potential for discovering and exploiting a more sophisticated machine learning model for improving the accuracy of the steel quality prediction when data is accumulated dramatically.

In the future, we will consider other model combination strategies together with other types of base learners, such as neural networks with the fuzzy system [12] [16]. Meanwhile, with the massive data sources, some machine learning strategies, such as deep learning, will be incorporated in the next stage. Notably, using deep neural networks can not only have the potential for mining latent information, but also relieving the workload of feature engineering.

References

[1] H. Shigemori and IAENG, "Quality and Operation Management System for Steel Products through Multivariate Statistical Process Control," in *Proceedings of the World Congress on Engineering*, 2013, vol. 1, pp. 119-124: IAENG London.

[2] R. Godina, E. M. Rodrigues, and J. C. Matias, "An Alternative Test of Normality for Improving SPC in a Portuguese Automotive SME," in *Closing the Gap Between Practice and Research in Industrial Engineering: Springer*, 2018, pp. 277-285.

[3] B. Lennox, G. Montague, and O. Marjanovic, "Detection of faults in Batch Processes: Application to an industrial fermentation and a steel making process," *Water Science and Technology*, 2000.

[4] S. Wang, J. Wan, D. Zhang, D. Li, and C. Zhang, "Towards Smart Factory for Industry 4.0: a Self-organized Multi-agent System with Big Data Based Feedback and Coordination," *Computer Networks*, vol. 101, pp. 158-168, 2016.

[5] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben, "Machine Learning in Manufacturing: Advantages, Challenges, and Applications," *Production & Manufacturing Research*, vol. 4, no. 1, pp. 23-45, 2016.

[6] E. Alpaydin, *Introduction to Machine Learning*. MIT press, 2009.

[7] D.-S. Kwak and K.-J. Kim, "A Data Mining Approach Considering Missing Values for the Optimization of Semiconductor-Manufacturing

Processes," *Expert Systems with Applications*, vol. 39, no. 3, pp. 2590-2596, 2012.

[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *Communications of the ACM*, vol. 39, no. 11, pp. 27-34, 1996.

[9] Y. Elovici and D. Braha, "A Decision-theoretic Approach to Data Mining," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 33, no. 1, pp. 42-51, 2003.

[10] A. K. Choudhary, J. A. Harding, and M. K. Tiwari, "Data Mining in Manufacturing: a Review Based on the Kind of Knowledge," *Journal of Intelligent Manufacturing*, vol. 20, no. 5, p. 501, 2009.

[11] H. Shigemori, M. Kano, and S. Hasebe, "Optimum Quality Design System for Steel Products through Locally Weighted Regression Model," *Journal of Process Control*, vol. 21, no. 2, pp. 293-301, 2011.

[12] D. Laha, Y. Ren, and P. N. Suganthan, "Modeling of Steelmaking Process with Effective Machine Learning Techniques," *Expert systems with applications*, vol. 42, no. 10, pp. 4687-4696, 2015.

[13] A. J. Tallón-Ballesteros and J. C. Riquelme, "Data Cleansing Meets Feature Selection: A Supervised Machine Learning Approach," in *International Work-Conference on the Interplay Between Natural and Artificial Computation*, 2015, pp. 369-378: Springer.

[14] D. Montgomery, "Statistical Quality Control, a Modern Introduction, 2009," Chichester, West Sussex, UK: Johm Wiley & sons Google Scholar.

[15] R. Jiang, *Introduction to Quality and Reliability Engineering*. Springer, 2015.

[16] Y. Bai, Z. Sun, J. Deng, L. Li, J. Long, and C. Li, "Manufacturing Quality Prediction Using Intelligent Learning Approaches: A Comparative Study," *Sustainability*, vol. 10, no. 1, p. 85, 2017.

[17] M. S. El-Din, H. Rashed, and M. El-Khabeery, "Statistical Process Control Charts Applied to Steelmaking Quality Improvement," *Quality Technology & Quantitative Management*, vol. 3, no. 4, pp. 473-491, 2006.

[18] Z. Ge and Z. Song, Multivariate Statistical Process Control: Process Monitoring Methods and Applications. Springer Science & Business Media, 2012.

[19] F. Megahed and L. Jones-Farmer, "A Statistical Process Monitoring Perspective on "Big Data", *Frontiers in Statistical Quality Control*, ed: Springer, New York, 2013.

[20] A. Dhini and I. Surjandari, "Review on Some Multivariate Statistical Process Control Methods for Process monitoring," in *Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management, Kuala Lumpur*, 2016, pp. 754-759.

[21] E. Alexopoulos, "Introduction to Multivariate Regression Analysis," *Hippokratia,* vol. 14, no. Suppl 1, p. 23, 2010.

[22] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data Mining and Analytics in the Process Industry: the Role of Machine Learning," *IEEE Access*, vol. 5, pp. 20590-20616, 2017.

[23] A. Dixit, Ensemble Machine Learning: A Beginner's Guide that Combines Powerful Machine Learning Algorithms to Build Optimized Models. Packt Publishing Ltd, 2017.

[24] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *International Workshop on Multiple Classifier Systems*, 2000, pp. 1-15: Springer.

[25] Z.-H. Zhou, "Ensemble Learning," *Encyclopedia of Biometrics*, pp. 411-416, 2015.

[26] T. Ditterrich, "Machine Learning Research: Four Current Direction," *Artificial Intelligence Magzine*, vol. 4, pp. 97-136, 1997.

[27] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC, 2012.

[28] F. Güneş, R. Wolfinger, and P.-Y. Tan, "Stacked Ensemble Models for Improved Prediction Accuracy."

[29] Z. Ye and H. Li, "Based on Radial Basis Kernel Function of Support Vector Machines for Speaker Recognition," in *Image and Signal Processing (CISP), 2012 5th International Congress on*, 2012, pp. 1584-1587: IEEE.

[30] G. Ke et al., "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146-3154.

[31] J. H. Friedman, "Greedy Function Approximation: a Gradient Boosting Machine," *Annals of statistics*, pp. 1189-1232, 2001.

[32] A. Natekin and A. Knoll, "Gradient Boosting Machines, a Tutorial," *Frontiers in Neurorobotics*, vol. 7, p. 21, 2013.

[33] T. Chen and C. Guestrin, "Xgboost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794: ACM.

[34] T. Chen, T. He, and M. Benesty, "Xgboost: Extreme Gradient Boosting," R Package Version 0.4-2, pp. 1-4, 2015.

[35] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling Neural Networks: Many Could Be Better Than All," *Artificial intelligence*, vol. 137, no. 1-2, pp. 239-263, 2002.

[36] A. Krogh and J. Vedelsby, "Neural Network Ensembles, Cross Validation, and Active Learning," in *Advances in Neural Information Processing Systems*, 1995, pp. 231-238.