

Cement Clinker Quality Prediction Using Advanced Analytics: A study of Machine Learning and Deep Learning Methods

Iqbal Qasim, Habeeb Balogun, Wusu Godoyon and Hafiz Alaka

This is a copy of the author's accepted version of a paper subsequently published in the proceedings of the 6th International Conference on Electrical, Computer and Energy Technologies, Rome, Italy, 06 - 09 Jul 2026.

The final published version will be available online at:

<https://ieeexplore.ieee.org/Xplore/home.jsp>

© 2026 IEEE . This manuscript version is made available under the CC-BY 4.0 license

<https://creativecommons.org/licenses/by/4.0/>

For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) license to any Accepted Manuscript version arising.

The WestminsterResearch online digital archive at the University of Westminster aims to make the research output of the University available to a wider audience. Copyright and Moral Rights remain with the authors and/or copyright owners.

Cement Clinker Quality Prediction Using Advanced Analytics: A study of Machine Learning and Deep Learning Methods

Iqbal Qasim*, Habeeb Balogun[†], Wusu Godoyon*, Hafiz Alaka*

*Big Data Technologies and Innovation Laboratory, University of Hertfordshire, Hatfield, UK

[†]School of Computer Science and Engineering, University of Westminster, London, UK

Email: I.qasim@herts.ac.uk, Baloguh@westminster.ac.uk, G.e.wusu@herts.ac.uk, H.alaka@herts.ac.uk

Abstract—Free lime content (f-CaO) in cement clinker is a crucial determinant of product quality in rotary kilns. Accurate real-time estimation of this parameter can significantly improve product quality and energy efficiency during production. While Machine Learning (ML) has shown promise for such predictive tasks, previous studies have primarily focused on predicting free lime content as a percentage, often resulting in high mean squared errors (MSE) and low R^2 values, making these approaches impractical for cement plants. This study developed and evaluated eight regression models, including a deep neural network model. The highest R^2 value of 0.39 was achieved using Support Vector Regression (SVR), while the lowest MSE of 0.08 was obtained with the deep neural network. However, these regression models did not meet the required confidence levels, prompting a shift to classification models. Subsequently, ten classification-based ML models were developed, evaluated, and compared, with Support Vector Machine with Radial Kernel (SVM R) emerging as the best performer, achieving an accuracy of 0.85 and an AUC of 0.91. These findings suggest that a robust classification-based ML model can be developed for predicting cement clinker quality.

Index Terms—Free lime content, Clinker quality, Predictive machine learning, Energy optimisation, Classification-based machine learning

I. INTRODUCTION

Cement clinker in a rotary kiln is the fundamental material used in cement production [1], derived from a precisely balanced raw meal with optimal granulometry, typically maintained below 2% [2]. The quality of cement clinker is influenced by several critical parameters that determine the cement's final properties and its performance in various applications. Among these, the free lime content is particularly crucial [3]. The acceptable range for free lime content in cement clinker is typically between 0.5% to 1.5% [4] or, according to some sources, below 2% [5]. A free lime content below 0.5% indicates over-burning, leading to unnecessary thermal energy consumption during production and a harder clinker, complicating the subsequent grinding process [3]. Conversely, a high free lime content suggests under-burning, which can lead to excessive expansion, affect clinker strength and stability, and cause potential cracking in construction [5]. Therefore, maintaining the free lime content within an acceptable range is essential for ensuring the final product's quality and the production process's energy efficiency.

In cement production, the detection of free lime content is often delayed due to the reliance on physical laboratory testing conducted after collecting clinker samples at the clinker cooler exit [3]. Fluctuations in free lime content can lead to unstable cement quality and increased energy consumption. Consequently, predicting free lime content in real time during production is crucial for implementing effective control strategies, stabilising cement quality, and reducing energy use [6]. Machine Learning (ML) has emerged as the most effective approach for accurate prediction tasks [7], [8]. As a result, significant efforts are being made to develop real-time prediction models for free lime content using machine learning algorithms trained on historical data [9].

In the past decade, numerous machine-learning (ML) algorithms have been presented to predict free lime content in cement clinker [4], [6], [10]–[14]. The algorithms widely used are Artificial Neural Networks (ANN), Gradient Boosting (GB), and Support Vector Machines (SVM). These models have typically been trained and evaluated on datasets with fewer than 5,000 samples [4], [10]. It is well-known that larger datasets generally lead to more accurate predictions. Still, the limited size of these datasets has often prevented complete model optimisation, sometimes resulting in less accurate predictions [15], [16]. In addition, the quality and selection of data and algorithms affect accuracy. [17]. Moreover, current state-of-the-art models predominantly rely on regression-based machine learning techniques, which may need to be fully optimised due to the complexity of the cement production process and the inherent ambiguity of datasets containing more than 100 controllable and uncontrollable variables.

Previous research has primarily focused on training regression-based models to predict continuous target variables, such as the exact value of free lime content in cement clinker. This approach is challenging due to the high precision required, where even minor errors can significantly affect the model's usefulness. In contrast, classification-based models can provide more interpretable results, especially for decision-making. Instead of relying on precise numerical predictions, stakeholders can work with clear categories (e.g., 'Good Quality Clinker' or 'Underburn Clinker') in the cement production process, directly related to quality control measures. We introduced a categorical target variable with two categories: 'Good

Quality Clinker' and 'Under-burn Clinker.' This adjustment makes our dataset suitable for training classification models. By categorising the outcomes, we generalise the prediction task, making it easier for the model to achieve higher accuracy in practical scenarios. Our approach shifts the focus from predicting exact values to determining whether the outcome meets a certain standard, which is more actionable in production settings. This shift is particularly beneficial because the precise free lime content is less important than whether it falls within an acceptable range (e.g., 0.5% to 1.5% for 'Good Quality Clinker'). In this research, we focused on the following key objectives:

- To develop and evaluate various classification-based machine learning algorithms
- To investigate the effect of feature selection on model performance
- To explore the impact of data size on model performance

The remainder of this paper is structured as follows: Section 2 contains the methodology of this work, including dataset description, data preprocessing, model development, and performance evaluation. In Section 3, the results and analysis are presented. Finally, in Section 4, the study is concluded, and its potential directions are discussed.

II. METHODOLOGY

Given the nature of the data and the focus on cement clinker quality, this study primarily employs regression-based supervised learning to predict the continuous value of free lime content in cement clinker. However, to meet the quality assurance requirements of the cement production plant, the dataset was also adapted for classification-based supervised learning. This approach allows for the assessment of clinker quality by categorising it as either 'Good Quality Clinker' or 'Underburn Clinker' based on the percentage of free lime present.

Our approach employs 18 machine learning techniques: eight are regression-based techniques to predict the amount of free lime content and ten for classifying the cement clinker quality. Techniques such as Linear Regression (LR) [18], Random Forest Regressor (RF) [19], [20], and Deep Neural Networks (DNN) have been utilised to predict various quality parameters in cement clinker, including free lime content, silica content, alumina content, sulphur trioxide content, and magnesium oxide content. These algorithms are well-regarded in cement clinker quality prediction and have been applied to diverse datasets. However, this study uniquely applies and compares these algorithms using a large dataset.

In our work, all the models were implemented using Python within a Sublime Text Integrated Development Environment (IDE). The experiments were conducted on an Apple MacBook Pro with macOS Ventura 13.0, equipped with an Apple M2 chip, 16 GB RAM, an 8-core CPU, and a 10-core GPU. The overall workflow of the proposed prediction system is illustrated in Fig. 1.

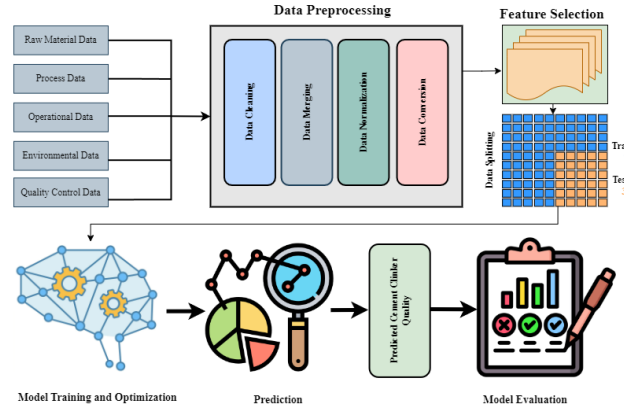


Fig. 1. Flowchart diagram of the prediction system.

A. Data Collection

The model was developed based on five types of data. This data includes raw material data, process data (preheater, kiln, cooling), operational data (production rate, energy consumption, operational parameters), environmental data, and quality control data (free lime content, silica ratio, alumina ratio, lime saturation factor). The dataset was collected from an operational cement plant in Indonesia. Each record in the dataset corresponds to an hourly operation and spans from 26th May 2022 to 31st Dec 2023, containing 14025 records and 116 variables (See appendix).

B. Data preprocessing

Data preprocessing is the first step of training an ML model. This step can be time-consuming and computationally expensive [21]; however, it is crucial for identifying and correcting invalid or inconsistent data for a smooth analysis process [22].

Data merging: we applied data merging to combine the operational, process, environmental, and quality control datasets. Data merging in this work is based on common variables (date and time) to map data across all datasets for a given date and time and to store a single record for each hour. The total number of records increased to 14025 after merging the datasets.

Data cleaning: is utilised to remove data that is ambiguous, to make the data consistent, and to complete missing values in the data. In this work, the merged dataset had incomplete data, which we addressed using median imputation [23]. However, we removed 1505 records with missing values in the target variable to avoid uncertainty and complications during model training and testing.

Data conversion: The merged dataset comprised all variables with continuous data, such as 'C5B Gas Temp °C', 'Fuel Main Burner tph', and 'Free lime percentage', among others. Initially, the dataset is suitable for regression ML algorithms to predict the free lime content in cement clinker. However, we introduced the 'Category' variable as a target variable. The

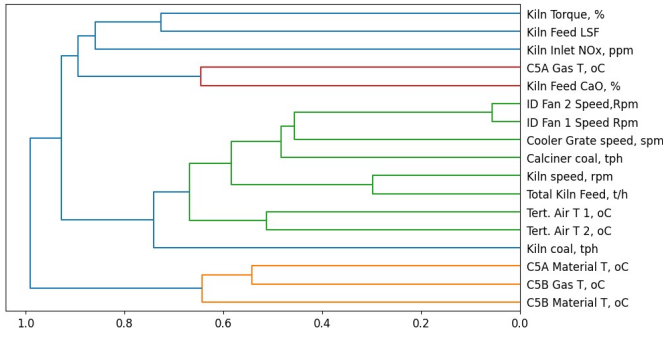


Fig. 2. Identification of redundant features.

values are allocated to the ‘Category’ variable [e.g., Good Quality Clinker = if the free lime content in percentage is between 0.5 and 1.5, while Underburn = if the value of free lime content is greater than 1.5] to have appropriate data for training classifier ML.

Data normalisation: is aimed at eliminating the influence of differing scales among features, as these often have unrelated dimensions [24]. Normalisation adjusts individual samples to unit norm, ensuring consistent scale and preventing issues during model implementation. The normalisation process applies the following formula to scale the dataset to values in the range 0-1.

$$X_{\text{Scaled}} = \frac{X - \text{median}(X)}{\text{IQR}(X)} \quad (1)$$

where:

- X is the original feature value.
- $\text{median}(X)$ is the median of the feature values.
- $\text{IQR}(X)$ is the interquartile range of the feature values, calculated as the difference between the 75th percentile and 25th percentile.

C. Feature Selection

Feature selection involves selecting features that affect the complexity and performance of the ML model. The feature selection step is mostly used to identify relevant features [25]. Currently, the final merged dataset contains 116 variables/features, out of which 35 were selected and these include ‘Kiln Feed HM’, ‘Fuel Main Burner t/h’, ‘Total Kiln Feed t/h’, ‘Clinker Temperature C° ’, ‘Grate Speed spam’, ‘CaO %’, ‘C5B Material Temp C° ’, ‘Kiln Cooler Energy kwh’ among many others. These conclusion was reached using different feature selection methods [26]. In addition, we illustrated these features relevance, see Figure 2 and Figure 3.

D. Model Development

The literature shows that predicting the absolute free lime value has yielded mixed results, with evaluation matrix values varying. Therefore, this research confirmed with domain experts that real-time prediction of free lime quality as ‘Good Quality Clinker’ or ‘Underburned Clinker’ would be highly accurate. This study utilised a supervised ML-based regression to

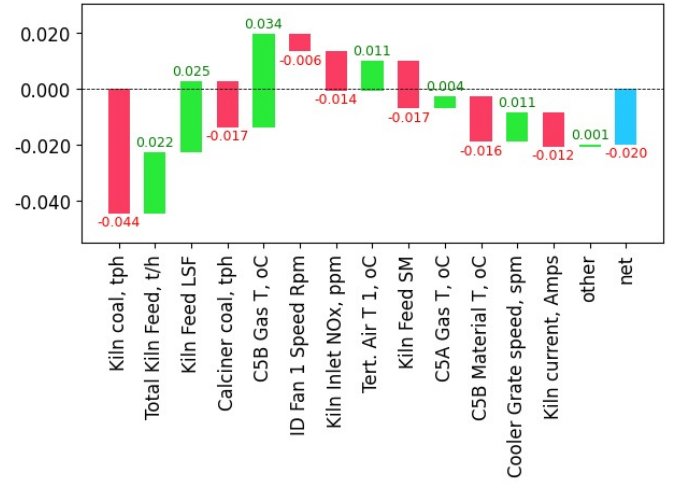


Fig. 3. Waterfall plot showing the positive and negative contribution of features to the dependent variable.

predict free lime content in cement clinker. Subsequently, this study develops a supervised machine learning approach based on classification models to predict cement clinker quality as ‘Good Quality Clinker’ or ‘Under-burn Clinker’ based on the percentage of free lime. To train the learning algorithms, we split the data into 70/30 training/test. The models developed and evaluated were using the default parameter configuration from Scikit-learn, except for a few whose parameters were adjusted. They include the DNN model, which is based on the feed-forward neural network (FFN) and implemented using the *Fastai*¹ library. This model comprises two hidden layers: the first with 650 neurons and the second with 375. Adding these two layers, along with input and output layers, results in a four-layer model. We used the *Fastai* default activation function RELU and the ‘Adam’ optimiser during hyperparameter setting. In addition, we implemented SVM with the following parameter settings: ‘C’ = 1.0, epsilon = 0.1, and a radial kernel. Furthermore, the RF was configured with 100 estimators, and the value of ‘K’ in the K-NN implementation was set to 5 neighbours with uniform weights and a leaf size of 30. Likewise, gradient boosting was implemented with a learning rate of 0.1, 100 estimators, and a least-squares regression loss function.

E. Evaluation of the Model

Initially, our dataset is feasible for regression modelling; however, by introducing a new target variable, ‘Category’ (i.e., if the percentage of free lime content is between 0.5 and 1.5, then it will be considered as ‘Good Quality Clinker,’ and if the percentage is greater than 1.5, then it will be regarded as ‘Underburned Clinker’), we make it feasible for binary classification as well. Therefore, here we introduce performance measures for both types of modelling. We used R-squared R^2 , Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to evaluate all the

¹<https://docs.fast.ai/>

regression-based prediction models. The metrics such as MSE and RMSE have been widely utilized in the prediction of cement clinker quality [14].

Predictions fall into four categories for binary classification, such as cement clinker quality as ‘Good Quality Clinker’ (1) or ‘Underburn Clinker’ (0). It includes true positive(TP), where the number of instances correctly predicted as ‘Good Quality Clinker’; false positives(FP), where the number of instances incorrectly predicted as ‘Good Quality Clinker’ when they are actually ‘Underburn’; false negative(FN), where the number of instances incorrectly predicted as ‘Underburn’ when they are actually ‘Good Quality Clinker’; and true negatives(TN), where the number of instances are correctly predicted as ‘Underburn’. Readers interested in these metrics can check [32].

Furthermore, the resampling technique is commonly used to evaluate ML models and estimate how well they will generalise to test data. K-fold cross-validation is a common and efficient method and was employed in this paper. In K-fold cross-validation (KfCV), the original dataset is randomly divided into K folds, each containing roughly equal observations. One of the K folds is set aside as the testing set, while the remaining k-1 folds are used for training the model. The model’s performance is then assessed using the held-out set, repeating this process k times, with each fold serving as the validation set. Performance metrics were recorded for each fold, and the model’s overall performance is evaluated by averaging across the k-folds. Typical values are $k = 5$ and $k = 10$; we used $k = 5$ in this study.

III. RESULT AND DISCUSSION

The analysis reveals significant differences in the predictive models’ results. The effectiveness of the models in this study can be determined by those values of MAE, MSE, and RMSE that are closer to zero. Additionally, the model performs better when R^2 values are closer to 1. Based on the analysis of the error rate, the DNN was identified as the most effective model for predicting free lime content in cement clinker. The SVR-based model outperforms the XGBoost-based model in terms of R^2 predicting cement clinker quality, ranking second among the predictive models according to the evaluation criteria. DNN, ANN, GB, RF, and SVR-based models are directly comparable, as are stacking and LR. DNN takes longer to train; however, it performs slightly better than XGBoost. Finally, the LR model takes less time during training but performs worse than other models. Values in bold in Table I represent the best performance.

The box plot method was used to illustrate the model’s performance, as shown in Figure 4. The plots in this figure are drawn based on 35 selected features. DNN was the most effective model across MAE, MSE, and RMSE, and visualised performance metrics were implemented. However, SVR yields a higher R^2 value than other predictive models. Additionally, the analysis shows that stacking and LR had the lowest performance, as they are closest to 0.0 for R^2 based on the interquartile range outliers and have the highest MAE, MSE,

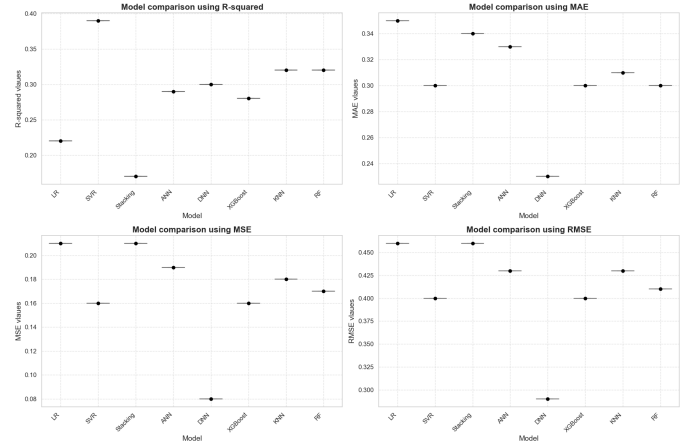


Fig. 4. Model comparison using R^2 , MAE, MSE, and RMSE shown as box plots.

and RMSE values in the box plots. Furthermore, models such as stacking, LR, ANN, and K-NN were the least accurate compared to DNN, SVR, and XGBoost across all performance measures. Therefore, Figure 4 and Table I show that the DNN model is the best predictor of free lime content in cement clinker, followed by SVR, XGBoost, and RF, which also produced good results.

To assess the impact of data size on model performance, all models were developed and evaluated using two datasets: one with 5,507 instances and the other with 10,800 instances. These instances were extracted after data preprocessing, and the results are summarised in Table I below. The findings indicate that the Deep Neural Network (DNN) consistently emerged as the most effective model for predicting cement clinker quality across both dataset sizes. In contrast, Linear Regression (LR) was identified as the most computationally efficient model. Notably, there was a performance difference between the two dataset sizes, with models trained on the larger dataset generally yielding better results. Increasing the dataset size to over 50,000 instances could further enhance model accuracy, supporting the notion that larger datasets yield more accurate outcomes [15], [16]. Additionally, the models developed with the larger dataset generally required more training time, particularly for DNN, ANN, and stacking models. The most efficient models, as measured by the performance metrics, are highlighted in bold.

Initially, the final dataset was employed to train eight different regression algorithms to predict the percentage of free lime in cement clinker. While the results for evaluation metrics such as MAE, MSE, and RMSE were acceptable, the R^2 values across all models were notably low, with none exceeding 0.39. This low R^2 value indicates that the models may not be reliable for accurate prediction. Consequently, based on expert domain opinion, we restructured the dataset to treat the problem as a binary classification task. A new target variable, ‘Category’, was introduced, categorising the records into two classes: ‘Good Quality Clinker’ for free lime

TABLE I
PERFORMANCE OF MODELS USING TWO DATASET SIZES

Model	5,507 Instances		10,800 Instances	
	Time (s)	R^2	Time (s)	R^2
Deep Neural Network	5.44	0.30	8.16	0.30
Support Vector Regressor	0.18	0.28	0.40	0.39
Extreme Gradient Boosting	0.37	0.23	1.10	0.28
Artificial Neural Network	4.29	0.23	6.37	0.29
Linear Regression	0.0021	0.17	0.0034	0.22
Stacking	6.61	0.16	13.62	0.17
K Nearest Neighbour	0.02	0.21	0.02	0.32
Random Forest	3.60	0.28	3.60	0.32

values between 0.5 and 1.5 and ‘Underburn Clinker’ for values exceeding 1.5.

Table II and III present the performance of ten widely recognised classification algorithms, chosen for their common use and efficiency [27]–[30], across two feature sets: one utilising all available features and another using features selected through feature selection (FS). The table includes metrics generated from a 5-fold cross-validation (CV) applied to a 75% balanced training set, with the models tested on the remaining 25% of untouched data.

Tables II and III offer a comprehensive comparison of model performance across various evaluation metrics—Accuracy (Acc), Precision (Pre), F1 score (F1), Recall (Rec), and Area Under the Curve (AUC)—for different classification algorithms. These metrics were calculated through 5-fold cross-validation and then tested on the untouched, imbalanced data (25% of the total dataset). Tables II and III correspond to the metrics calculated for the different feature sets: all features and features selected through the FS technique, respectively. Each table compares the metrics obtained from the CV phase with those from the testing phase. This comparison helps assess whether the model’s performance on the training data indicates its generalizability to unseen data. If the performance metrics from cross-validation exceed those from testing, it suggests potential overfitting, where the model learns to memorise the training data instead of learning underlying patterns. Conversely, suppose the performance metrics from testing are comparable to or better than those from cross-validation. In that case, the model is not overfitting and can generalise well to new, unseen data. In Tables II and III, all models demonstrate consistent performance across both phases, suggesting robust generalisation. Based on performance metrics in Tables II and III, the best machine learning models can be identified by examining AUC, accuracy, precision, recall, and F1 scores.

From Table II, the Support Vector Machine with Polynomial Kernel (SVM_P) consistently performs across cross-validation and testing phases. It achieves an AUC of 0.91 and AUC_V of 0.90, with an accuracy of 0.83 in both phases and balanced precision, recall, and F1 scores, indicating a robust and generalisable model. Similarly, the Support Vector Machine with Radial Kernel (SVM_R) also performs well, with an AUC of 0.90 for both cross-validation and testing, an accuracy of 0.83

TABLE II
CLASSIFICATION PERFORMANCE USING ALL FEATURES

Model	Acc	Acc _V	Pre	Pre _V	Rec	Rec _V	F1	F1 _V	AUC	AUC _V
GB	0.82	0.79	0.82	0.80	0.83	0.78	0.82	0.79	0.88	0.88
MLP	0.83	0.78	0.84	0.76	0.83	0.81	0.83	0.78	0.88	0.86
RF	0.80	0.79	0.82	0.82	0.78	0.76	0.80	0.79	0.87	0.88
SVM _L	0.72	0.74	0.71	0.74	0.74	0.73	0.72	0.74	0.79	0.79
SVM _P	0.83	0.83	0.85	0.85	0.80	0.82	0.83	0.83	0.91	0.90
SVM _R	0.83	0.83	0.85	0.85	0.82	0.82	0.83	0.83	0.90	0.90
LR	0.75	0.75	0.72	0.76	0.80	0.75	0.76	0.75	0.80	0.81
DA	0.71	0.71	0.69	0.73	0.77	0.69	0.73	0.71	0.78	0.78
KNN	0.74	0.70	0.71	0.67	0.82	0.78	0.76	0.72	0.82	0.77
DT	0.70	0.72	0.72	0.71	0.67	0.76	0.69	0.73	0.77	0.72

($\cdot$)_V denotes validation performance.

in both phases, and balanced precision, recall, and F1 scores, showcasing strong generalisation capabilities. The Gradient Boosting (GB) model is also noteworthy, achieving an AUC of 0.88 across both phases, an accuracy of 0.82 during cross-validation and 0.79 during testing, with metrics suggesting reliable performance on both balanced and imbalanced datasets.

From Table III, Gradient Boosting (GB) continues to excel, achieving an AUC of 0.93 in cross-validation and 0.92 in testing, along with an accuracy of 0.85 and 0.84, respectively, and high scores across precision, recall, and F1. The Support Vector Machine with Radial Kernel (SVM_R) also maintains strong performance with an AUC of 0.92 in both phases, accuracy scores of 0.84 and 0.85, and balanced metrics, indicating its effectiveness across different feature sets. Additionally, the Random Forest (RF) model shows solid performance on the dataset with selected features, achieving an AUC of 0.92 in cross-validation and 0.91 in testing and an accuracy of 0.84 in both phases, with balanced metrics, making it a reliable choice.

Overall, the top three models, based on their consistent and robust performance across all metrics, were Gradient Boosting (GB), Support Vector Machine with Radial Kernel (SVM_R), and Random Forest (RF) with feature selection (FS). These models demonstrated strong generalisation capabilities without overfitting, making them ideal candidates for deployment. Table IV presents the metrics generated from the test set for these top three models and the features used. The SVM with radial kernel is an exceptionally robust choice due to its well-rounded performance across multiple evaluation criteria. With an AUC of 91%, this model can distinguish between the two cement clinker classes, indicating high overall predictive effectiveness. The SVM’s F1 score of 85% demonstrates a balance between precision and recall, which is crucial in scenarios with severe consequences for false positives and false negatives. Furthermore, with an accuracy of 85%, the model demonstrates reliable performance across the entire dataset, making it a dependable choice for general use.

IV. CONCLUSION

This study compares eight machine learning techniques: DNN, ANN, SVR, DT, GB, LR, Stacking, RF, and KNN—for

TABLE III
CLASSIFICATION PERFORMANCE USING SELECTED FEATURES

Model	Acc	Acc _V	Pre	Pre _V	Rec	Rec _V	F1	F1 _V	AUC	AUC _V
GB	0.85	0.84	0.83	0.81	0.89	0.87	0.86	0.84	0.93	0.92
MLP	0.82	0.82	0.82	0.80	0.82	0.84	0.82	0.82	0.89	0.89
RF	0.84	0.84	0.83	0.81	0.90	0.87	0.86	0.83	0.92	0.91
SVM _L	0.72	0.72	0.71	0.70	0.74	0.77	0.73	0.73	0.78	0.78
SVM _P	0.82	0.83	0.80	0.81	0.84	0.85	0.82	0.83	0.87	0.88
SVM _R	0.84	0.85	0.83	0.83	0.85	0.87	0.84	0.85	0.92	0.91
LR	0.72	0.72	0.71	0.69	0.75	0.78	0.73	0.73	0.78	0.78
DA	0.73	0.73	0.71	0.70	0.76	0.79	0.74	0.74	0.78	0.78
KNN	0.79	0.79	0.73	0.73	0.91	0.91	0.81	0.81	0.87	0.87
DT	0.76	0.76	0.74	0.74	0.80	0.79	0.77	0.77	0.76	0.76

(·)_V denotes validation performance.

TABLE IV
PERFORMANCE OF THE TOP THREE MODELS

Model	Features	Acc	Pre	Rec	F1	AUC
GB	FS features	0.84	0.81	0.87	0.84	0.92
SVM _R	FS features	0.85	0.83	0.87	0.85	0.91
RF	FS features	0.84	0.81	0.87	0.84	0.91

predicting free lime content in cement clinker. Among these, a DNN implemented with the Fastai API outperformed the others, achieving the lowest RMSE in predicting free lime content. Nevertheless, other models, such as SVR and GB, also showed strong predictive performance, with SVR achieving a high R^2 value. Regarding computational efficiency, LR was the fastest, with a training time of just 0.0034 seconds. Additionally, the study examined the impact of dataset size on model performance, revealing that larger datasets significantly enhance predictive accuracy.

Despite these advancements, all regression-based ML models in this study struggled to achieve high accuracy in predicting absolute free lime content. In contrast, classification-based models proved more effective, offering higher accuracy and providing more actionable insights for domain experts.

The study concludes that classification-based ML models are more practical and effective than regression-based models for predicting free lime content in cement clinker. This study explored 10 classification models to predict clinker quality, focusing on developing explainable, generalisable models. A comprehensive analysis of 116 variables identified 35 key features using FS, which, when used in models such as GB and SVM with an RBF kernel, yielded the most accurate predictions. Furthermore, models utilising the 35 selected variables outperformed those using the full set of 116 variables, underscoring the importance of feature selection for cement clinker quality prediction.

The findings of this paper highlight the relevance of the selected variables and the robustness of the ML model development and validation approach. This research contributes to academic knowledge and offers practical guidance for industry experts in the cement sector. It provides a foundation for future research exploring AI and ML applications in cement quality

prediction.

REFERENCES

- [1] K. Kaewmanee *et al.*, "Effect of free lime content on properties of cement-fly ash mixtures," *Construction and Building Materials*, vol. 38, pp. 829–836, 2013.
- [2] A. Loukili *et al.*, "A new approach to determine autogenous shrinkage of mortar at an early age considering temperature history," *Cement and Concrete Research*, vol. 30, no. 6, pp. 915–922, 2000.
- [3] F. Alemayehu and O. Sahu, "Minimization of variation in clinker quality," *Advances in Materials*, vol. 2, no. 2, pp. 23–28, 2013.
- [4] X. Liu *et al.*, "A novel support vector machine ensemble model for estimation of free lime content in cement clinkers," *ISA Transactions*, vol. 99, pp. 479–487, 2020.
- [5] N. H. Mtarfi, Z. Rais, and M. Taleb, "Effect of clinker free lime and cement fineness on the cement physicochemical properties," *J. Mater. Environ. Sci.*, vol. 8, no. 7, pp. 2541–2548, 2017.
- [6] X. Hao *et al.*, "Prediction of f-cao content in cement clinker: A novel prediction method based on lightgbm and bayesian optimization," *Chemometrics and Intelligent Laboratory Systems*, vol. 220, p. 104461, 2022.
- [7] Y. Vorobeychik and J. R. Wallrabenstein, "Using machine learning for operational decisions in adversarial environments," Sandia National Lab. (SNL-CA), Livermore, CA (United States), Tech. Rep. SAND2013-1328C, 2013.
- [8] V. Ríos Canales, "Using a supervised learning model: Two-class boosted decision tree algorithm for income prediction," *Computer Engineering*, 2016.
- [9] K. Boakye, K. Fenton, and S. Simske, "Machine learning algorithm to predict co2 using a cement manufacturing historic production variables dataset: A case study at union bridge plant, heidelberg materials, maryland," *Journal of Manufacturing and Materials Processing*, vol. 7, no. 6, p. 199, 2023.
- [10] X. Lyu *et al.*, "Quality control of cement clinker through operating condition classification and free calcium oxide content prediction," *Applied Sciences*, vol. 14, no. 3, p. 1119, 2024.
- [11] J. Yuan *et al.*, "Prediction of free lime content in cement clinker based on rbf neural network," *Journal of Wuhan University of Technology-Mater. Sci. Ed.*, vol. 27, no. 1, pp. 187–190, 2012.
- [12] P. Zhao, Z. Zhao, and Y. Chen, "An improved combination modeling method for free lime content in cement clinker," in *2020 Chinese Automation Congress (CAC)*. IEEE, 2020, pp. 1–6.
- [13] W. Li, D. Wang, and T. Chai, "Multisource data ensemble modeling for clinker free lime content estimate in rotary kiln sintering processes," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 2, pp. 303–314, 2014.
- [14] A. M. Ali, J. D. Tabares, and M. W. McGinley, "A machine learning approach for clinker quality prediction and nonlinear model predictive control design for a rotary cement kiln," *Journal of Advanced Manufacturing and Processing*, vol. 4, no. 4, p. e10137, 2022.
- [15] S. Lee *et al.*, "Data mining-based predictive model to determine project financial success using project definition parameters," in *Proceedings of the 28th International Symposium on Automation and Robotics in Construction, ISARC, Seoul, Korea*, vol. 29, 2011, pp. 123–130.
- [16] K. Kaur and O. P. Gupta, "A machine learning approach to determine maturity stages of tomatoes," *Oriental Journal of Computer Science and Technology*, vol. 10, no. 3, pp. 683–690, 2017.
- [17] J. Runge and R. Zmeureanu, "Forecasting energy use in buildings using artificial neural networks: A review," *Energies*, vol. 12, no. 17, p. 3254, 2019.
- [18] A. M. Al-Swaidani *et al.*, "Development of multiple linear regression, artificial neural networks and fuzzy logic models to predict the efficiency factor and durability indicator of nano natural pozzolana as cement additive," *Journal of Building Engineering*, vol. 52, p. 104475, 2022.
- [19] H. Zermane and A. Drardja, "Development of an efficient cement production monitoring system based on the improved random forest algorithm," *The International Journal of Advanced Manufacturing Technology*, vol. 120, no. 3, pp. 1853–1866, 2022.
- [20] Z. Yin *et al.*, "Research on cement granularity prediction based on random forest," in *2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC)*, vol. 6. IEEE, 2022, pp. 0–0.

- [21] M. K. M. Shapi, N. A. Ramli, and L. J. Awalin, "Energy consumption prediction by using machine learning for smart building: Case study in malaysia," *Developments in the Built Environment*, vol. 5, p. 100037, 2021.
- [22] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1192–1205, 2018.
- [23] W.-C. Lin and C.-F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," *Artificial Intelligence Review*, vol. 53, pp. 1487–1509, 2020.
- [24] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, p. 105524, 2020.
- [25] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, pp. 1157–1182, 2003.
- [26] H. Balogun, H. Alaka, C. Egwim, and S. O. Ajayi, "Random forest feature selection for particulate matter (pm10) pollution concentration," pp. 576–587, 2021.
- [27] R. Olu-Ajayi, "An investigation into the suitability of k-nearest neighbour (k-nn) for software effort estimation," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, 2017.
- [28] R. Olu-Ajayi *et al.*, "Machine learning for energy performance prediction at the design stage of buildings," *Energy for Sustainable Development*, vol. 66, pp. 12–25, 2022.
- [29] H. Balogun *et al.*, "Boruta-grid-search least square support vector machine for no2 pollution prediction using big data analytics and iot emission sensors," *Applied Computing and Informatics*, vol. 21, pp. 101–113, 2021.
- [30] H. Balogun, H. Alaka, E. Demir, C. N. Egwim, G. E. Wusu, W. Yusuf, M. Adegoke, and I. Qasim, "Deconstructability prediction for building using machine learning and ensemble feature selection techniques," *Scientific Reports*, vol. 15, p. 22152, 2025.