Lung Cancer Prediction from Smoking Cause by Machine Learning Classification Models

Nada M. Sallam [1] *, **P. K. Dutta** [2]
[1] Faculty of Computer Studies, Arab Open University, Riyadh, 11681, Saudi Arabia.
[2] School of Engineering and Technology, Amity University Kolkata, India.

Emails: n.sallam@arabou.edu.sa, pkdutta@kol.amity.edu

Abstract

The incidence of lung cancer varies in males and females, which occurs due to the abnormal and uncontrolled growth of cells in the lungs. It has a greater predilection in males as compared to females. Smoking is the most important risk factor for lung cancer. It causes serious breathing issues and also affects other organs. It increases the mortality rate both in young adults as well as in the older age group. Therefore, there is improvement in medical technologies to facilitate specialized diagnosis and treatment, but the mortality has not been controlled to a satisfactory extent. It is important to take preventive measures and precautions at the initial stages. Machine learning brings various advancements to the medical sector due to which various diseases can be detected at an early stage. In this paper, we presented different machine learning classifier techniques used for the classification of the present lung cancer data in the UCI machine learning repository as benign and malignant. The dataset is divided into cancerous and non-cancerous by converting the input data into binary form and using the classifier technique in the Weka tool. This specifically includes classifiers used: Logistic Regression, Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Decision Trees, and Naïve Bayes. In addition, we study the effect of data preprocessing methods on our prediction accuracy, such as data normalization and feature selection. The study seeks to help develop various reliable resources for lung cancer identification, which are critical for diagnosing and treating patients in a timely manner and improving their outcomes.

Keywords: Lung cancer prediction; Machine learning classifiers; Smoking risk factors; Medical data preprocessing; Early cancer diagnosis.

1 Introduction

Artificial Intelligence, or specifically termed as Machine Learning (ML), has taken a frontline position in availing medical practice with efficient and effective tools for improved diagnosis and treatment of diseases such as lung cancer [1]. A key feature of ML algorithms is their capability to solve complex data analyses within relatively short periods and highlight descriptive features important for making timely diagnoses [1]. For instance, algorithms identify cancer kernels and tumors in visual anatomy such as CAT scans and X-rays to aid radiologists in accurate and early diagnosis [2]. This capability allows patient treatment to be determined with high accuracy and in a shorter time compared to traditional tools [2].

Moreover, ML contributes to the development of molecular markers for lung cancer patients. By integrating genotype and phenotype data, personalized algorithms effectively search for the most optimal treatment plans while excluding futile therapies [3]. This approach not only increases the efficacy of the treatment but also decreases dangerous side effects and costs, which traditional healthcare services cannot overcome [4].

https://doi.org/10.54216/JAIM.090105

Received: October 21, 2024 Revised: December 17, 2024 Accepted: January 19, 2025

Beyond diagnosis and therapy, ML plays an important role in risk assessment and prognosis prediction [5]. ML algorithms can identify an individual's predisposition to lung cancer using information derived from electronic health records (EHRs), genomic data, and lifestyle parameters [6]. This capability enables preventive screening activities focused on high-risk groups [7], thereby concentrating resources on early detection and treatment to prevent malignant tumors [8, 9].

ML also assists in drug identification and development. By analyzing big data, it is possible to predict the nature, efficiency, and dosage of particular drugs using ML models [10]. This reduces the time required for drug development, lowers costs for pharmaceutical companies, and provides exclusive treatments for specific cancer types [11].

However, challenges remain in the implementation of ML in lung cancer care. The quality of training data shapes ML model accuracy and reliability [12]. Bias in datasets leads to biased predictions and amplifies healthcare disparities [13]. Hence, addressing the problem of racially and ethnically imbalanced data is crucial for ensuring fair ML healthcare applications [13]. Furthermore, the explainability of ML model results, especially from deep learning algorithms, is a significant barrier to implementation. Clinicians and patients often struggle to understand the reasoning behind predictions, leading to mistrust [14]. Explainable AI (XAI) techniques are being developed to address this, enhancing confidence among healthcare professionals and patients by providing clear insights into predictions [15].

Emerging wearable health devices powered by ML are providing new options for patient monitoring [16]. These devices track physiological factors, respiratory patterns, and other features, offering relevant alerts for the treatment process [17]. These systems improve patient outcomes and reduce healthcare costs by facilitating early identification of risk factors and disease progression [18].

ML also integrates data from imaging, genetics, proteomics, and clinical records, enhancing the understanding of diseases and the development of treatments. Natural Language Processing (NLP) complements ML by mining text data such as clinical notes and research articles, enriching clinical practice and research [19]. However, risks like patient data privacy and minority data accuracy must be managed appropriately. Collaboration among researchers, clinicians, and policymakers is necessary to determine best practices for ML implementation in real-world scenarios.

Due to advancements in ML technologies, diagnosing and treating lung cancer has become more efficient. The analysis demonstrates the potential to diagnose lung cancer in its earliest stages and provide personalized treatment approaches, improving patients' quality of life.

2 Literature Review

3 Related Work

In this paper, we provide a summary of the most recent and relevant research works focused on predicting lung cancer occurrence using machine learning techniques and models.

The primary goal of the study [1] is to establish early detection of lung cancer by evaluating the performance of various classification algorithms. The authors used classification algorithms such as Support Vector Machine (SVM), Naive Bayes, Decision Tree, and Logistic Regression. Logistic Regression achieved an accuracy of 96.9% for the UCI lung cancer dataset, while SVM achieved a higher accuracy of 99.2%.

The study [2] aimed to enhance performance, measured by RMSE, and predict lung cancer survival time in terms of months ($\leq$ 6, 7–24, or $>$ 24). The Random Forest classification model was integrated with three regression models (Gradient-Boosted Machines and General Linear Regression). Random Forest was most effective for survival times $\leq$ 6 months (RMSE 10.52) and $>$ 24 months (RMSE 20.51), while Gradient-Boosted Machines were significant for survival times between 7–24 months (RMSE 15.65).

The research in [3] evaluates the efficacy of machine learning methods for lung cancer diagnosis using a dataset of 56 features and 32 samples. Preprocessing techniques such as normalization, dimensionality reduction,

https://doi.org/10.54216/JAIM.090105

Received: October 21, 2024 Revised: December 17, 2024 Accepted: January 19, 2025

and feature selection were applied before running six classification algorithms: Decision Tree (DT), Random Forest (RF), k-Nearest Neighbors (k-NN), Logistic Regression (LR), Support Vector Machines (SVM), and Naive Bayes (NB). k-NN achieved the highest accuracy with 83%, outperforming Naive Bayes (87%) and Decision Tree (71%).

The study [4] focused on applying machine learning to assess lung cancer risk and improve accuracy by incorporating effective algorithms. Techniques such as Random Forest, Logistic Regression, and XG-Boost were compared. The study used datasets from Cleveland hospital records and US public health databases (NLST and PLCO) containing demographic, lifestyle, and clinical information. The evaluation metrics included recall, accuracy, F1 Score, and precision.

In [5], machine learning was used to examine the effect of smoking on lung cells. Data from the Human Lung Cell Atlas included four cell types: endothelial, epithelial, immune, and stromal cells, along with smoking statuses: active smokers, former smokers, and nonsmokers. Eight machine learning algorithms ranked genes, identifying four potential biomarkers: *HLA-B*, *FTL*, *B2M*, and *HSP90B1*. A mathematical formula for DNA patterns concerning smoking was also developed.

The study [6] compares four machine learning models (SVM, DT, k-NN, and RF) for lung cancer diagnosis using datasets with corresponding parameters. The accuracy ranged from 98% to 100%. To improve interpretability, techniques such as Local Interpretable Model-agnostic Explanations (LIME) and decision boundaries were used.

The authors of [7] analyzed e-questionnaire data stratified by smoking status (never, ex, and current smokers) using stochastic gradient boosting (SGB). The models achieved accuracies of 82% for never smokers, 77% for current smokers, and 63% for former smokers. Significant variables included age, sex, education level, and shortness of breath during exercise.

The research [8] focused on predicting smoking cessation using models such as Logistic Regression (LoR), CART, ANN, NB, RF, k-NN, and SVM. ANN achieved the highest sensitivity (70.4%), specificity (56.7%), and accuracy (64%), providing general practitioners with reliable estimates for individualized therapies.

The work in [9] proposed a deep learning model for lung cancer prognosis using data from the Taipei Medical University Clinical Research Database and the Taiwan Cancer Registry. Using nine machine learning algorithms, ANN showed the highest accuracy with an AUC of 0.89. Significant biomarkers included cancer stage, age, smoking habits, and EGFR gene status.

Authors in [10] utilized a public dataset for lung cancer prognosis using 10-fold cross-validation. The Rotation Forest model achieved the highest accuracy of 97.1% and AUC of 99.3%.

Research in [11] proposed an ensemble classifier based on patient symptoms and risk factors. The ensemble model outperformed individual classifiers, demonstrating improved accuracy in lung cancer pre-diagnosis.

The study [12] explored interpretability methods for three models (BACH, PLCOm2012, and LCART) and used Explainable AI (xAI) techniques to highlight significant features. This enhanced model transparency and confidence in predictions.

In [13], researchers analyzed data from 1,000 patients with 23 features. Classifiers such as Naive Bayes, k-NN, Decision Tree, and Random Forest achieved 100% accuracy on the full dataset and 98.7% on a reduced feature set.

The work [14] tested classifiers (k-NN, Naive Bayes, RBF, and J48) on the UCI repository dataset. RBF was the most accurate, achieving 81.25% accuracy.

Finally, [15] compared Gradient Boosted Trees (GBT), Random Forest, Majority Voting, and Neural Network models. GBT achieved the highest accuracy of 90%, highlighting its potential for lung cancer diagnosis.

The following Table 1 summarizes the recent research related to lung cancer prediction.

https://doi.org/10.54216/JAIM.090105

Received: October 21, 2024 Revised: December 17, 2024 Accepted: January 19, 2025

Table 1: Summary of recent research on lung cancer prediction.

Model/Method	**Year**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**Advantages/Disadvantages**
RBF	-	81.25	-	-	Effective for prediction, but performance varies by implementation
Decision Tree	2022	99.2	-	-	Easy to interpret; prone to overfitting
Naive Bayes	2022	99.2	-	-	Efficient; assumes feature independence
SVM	2022	99.2	-	-	Handles high-dimensional data; computationally expensive
Logistic Regression	2022	99.2	-	-	Simple to implement; limited to linear relationships
Neural Network	2024	91.98	85.76	85.76	High accuracy; "black box" problem
GBT	2024	91.98	85.76	85.76	Outperforms individual models; computationally expensive
Random Forest	2022/2024	92.5/91.98	85.76/92.5	85.76/-	Robust to overfitting; harder to interpret
Majority Voting	2024	91.98	85.76	85.76	Combines classifiers for higher accuracy

4 Material and Methods

4.1 Data Acquisition

The data employed in this work is sourced from a public dataset on lung cancer, available through the UCI Machine Learning Repository. The dataset contains information about patient characteristics such as age, gender, smoking status, and other relevant variables. A link to the dataset is available on Kaggle [17].

Figure 1 shows the histogram of the variables within the tested dataset. Figure 2 presents the histogram of the age variable in more detail. Figure 3 provides a correlation analysis of the variables with respect to one another.

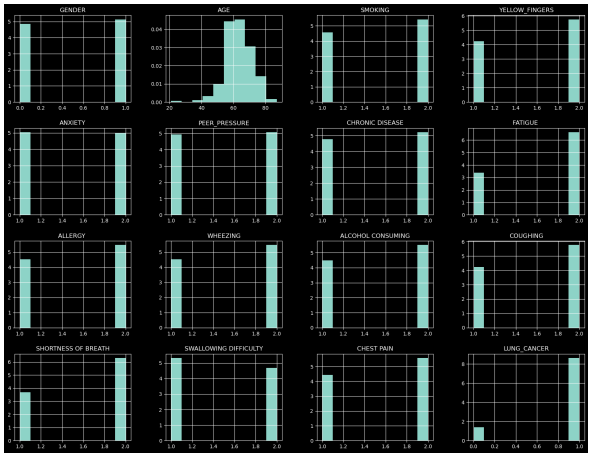

Figure 1: Histogram of the variables within the tested dataset.

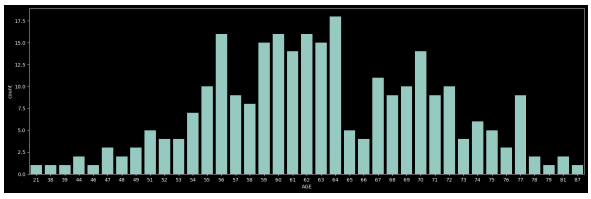

Figure 2: Histogram of the age variable within the tested dataset.

Received: October 21, 2024 Revised: December 17, 2024 Accepted: January 19, 2025

Figure 3: Correlation analysis of the variables within the tested dataset.

4.2 Data Preprocessing

Due to the presence of missing values in the dataset, data imputation techniques were applied. Alternatively, instances with significant missing values were filtered out. To ensure uniformity, all features were normalized or standardized to a comparable range. This step prevents large or small absolute values of specific features from influencing the model disproportionately.

In cases where the number of features is significantly larger than the number of instances, feature selection techniques were employed. These methods identify a subset of features containing the maximum entropy, likely increasing model accuracy while mitigating interpretability issues.

4.3 Model Selection

This section discusses the machine learning classification models selected for their known performance in medical diagnosis and prediction:

- **Support Vector Machines (SVMs):** SVMs are highly effective at identifying the best separation boundary (hyperplane) by focusing on the data points closest to the boundary (support vectors). This method maximizes the margin, making SVMs particularly strong for high-dimensional datasets. However, they can be computationally expensive for large datasets, especially those with numerous dimensions [18].
- **k-Nearest Neighbors (k-NN):** k-NN is a lazy learner that does not build a model but instead stores the training data. For each new data point, it identifies the ”k” nearest neighbors and determines the majority class. While k-NN is simple and easy to use, its performance is sensitive to the choice of ”k” and the distance metric. It is often applied in recommendation systems and anomaly detection [19].
- **Random Forest:** Random Forest uses an ensemble of decision trees trained on random subsets of the data and features. By averaging their predictions, Random Forest reduces overfitting and performs well in high-dimensional datasets. It is widely used in applications such as stock price prediction and medical diagnosis due to its precision and robustness [20].
- **Naive Bayes:** Despite its assumption of feature independence, Naive Bayes is an efficient classifier, particularly for text data. It uses Bayesian theorem to estimate the probability of a given data point belonging to a particular class. This simplicity makes it efficient for applications like spam filtering and document classification [21].
- **Logistic Regression:** Logistic Regression predicts the probability of a binary outcome using a logistic function. It is well-suited for applications where the probability of occurrence is critical, such as customer churn prediction and biomedical studies to identify disease-causing factors [22].
- **Decision Tree:** Decision Trees provide a transparent and interpretable way of making decisions, similar to a flowchart where each vertex represents a feature and each edge represents an outcome. They are commonly used in credit risk assessment and fraud detection. However, they are prone to overfitting, particularly with noisy data [23].

https://doi.org/10.54216/JAIM.090105

Received: October 21, 2024 Revised: December 17, 2024 Accepted: January 19, 2025

5 Experimental Results

5.1 Dataset Preparation

The dataset was divided into training and testing sets. Selected models were trained on the training set and evaluated on the test data using performance metrics such as recall, accuracy, F1-score, and precision [9]. Two key parameters considered were the number of iterations and the use of cross-validation techniques, such as k-fold cross-validation, to improve performance estimates.

The results were evaluated by comparing models based on chosen metrics, and the best-performing models were identified. Models were ranked by performance, interpretability, and computational efficiency. The impact of different preprocessing techniques on model accuracy was also explored. If imaging data, such as computed tomography (CT), were present in the dataset, deep learning models like Convolutional Neural Networks (CNNs) were employed for analysis. Model explainability was explored using techniques like LIME and SHAP to understand feature importance and build trust in the models' predictions.

5.2 Performance Metrics

The performance metrics used in this work are defined as follows:

- **Accuracy:** Measures overall correctness but may be misleading with imbalanced data.
- **F1-Score:** Balances precision and recall, especially important for imbalanced datasets.
- **Precision:** Focuses on avoiding false positives.
- **Recall (Sensitivity):** Focuses on capturing all true positives.
- **ROC-AUC (Receiver Operating Characteristic - Area Under Curve):** Evaluates the model's ability to discriminate between classes, providing a comprehensive measure of performance.

5.3 Results and Discussion

The performance results of different models for the tested dataset are shown in Table 2. Based on multiple metrics, the SGD Classifier demonstrated the best overall performance, offering high accuracy, F1-Score, and ROC-AUC, while maintaining strong recall. Logistic Regression was a close second.

Model	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	ROC-AUC (%)
Logistic Regression	92.86	95.65	100.0	91.67	95.83
Random Forest	87.50	92.63	100.0	86.27	93.14
Gaussian NB	91.07	94.51	97.73	91.49	90.19
K-Neighbors Classifier	83.93	90.53	97.73	84.31	82.16
SGD Classifier	94.64	96.70	100.0	93.62	96.81
Decision Tree	87.50	92.63	100.0	86.27	93.14
AdaBoost Classifier	85.71	91.67	100.0	84.62	92.31

Table 2: The output results of different models for the tested dataset.

Figure 4 shows the accuracy summary of different models for the tested dataset.

https://doi.org/10.54216/JAIM.090105

Received: October 21, 2024 Revised: December 17, 2024 Accepted: January 19, 2025

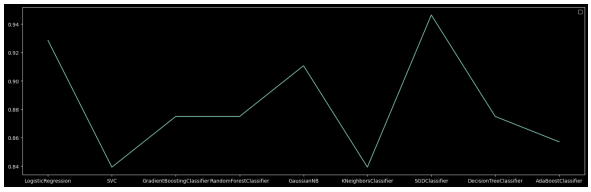

Figure 4: The output accuracy summary of different models for the tested dataset.

6 Conclusion and Discussion

Machine Learning (ML) has revolutionized the diagnosis, treatment, and management of lung cancer, demonstrating its potential as a transformative technology in the healthcare sector. The future of lung cancer research and treatment lies in the promising capabilities of ML, including its ability to analyze big data and derive meaningful insights.

The two most common applications of ML in lung cancer are early diagnosis and detection. By applying advanced ML algorithms to medical imagery, such as CT scans and X-rays, radiologists can accurately identify potentially cancerous nodules and tumors, enabling timely intervention and significantly improving patient outcomes. Additionally, ML plays a pivotal role in personalized medicine by leveraging patient-specific data, including genetic predispositions, lifestyle, and treatment history, to define optimal treatment plans and maximize therapeutic efficacy. This personalized approach not only improves treatment outcomes but also minimizes unnecessary side effects and healthcare costs.

Beyond diagnosis and treatment, ML extends to risk assessment and prognosis prediction. Algorithms can identify individuals at higher risk of developing lung cancer, enabling targeted screening and reducing risk factors. ML can also predict disease progression, providing valuable insights for treatment regimens and patient counseling.

Despite these advancements, challenges remain. The quality and nature of training datasets are critical for developing effective ML models. Biased data can lead to inaccurate predictions and exacerbate health inequalities. Therefore, diverse and unbiased datasets are essential for the equitable deployment of ML in healthcare.

Another challenge is the ”black box” nature of many ML models, particularly deep learning algorithms. Their lack of transparency complicates adoption by clinicians and patients. Explainable AI (XAI) techniques are being developed to address this issue, offering insights into model operations and enhancing trust in their predictions.

Ethical concerns, including data privacy, algorithm fairness, and the impact of automation on healthcare jobs, also need to be addressed. Strong regulatory frameworks are required to ensure the ethical and accountable use of ML technologies.

Overall, ML offers immense potential for lung cancer research and patient care. By leveraging its ability to process and analyze big data, ML provides opportunities for early diagnosis, personalized treatment, risk assessment, and prognosis prediction, ultimately improving patient outcomes and quality of life.

References

[1] R. Patra. Prediction of lung cancer using machine learning classifier. pages 132–142.

[2] M. I. Faisal, S. Bashir, S. Khan, Z. S. Khan, and F. H. Khan. An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. In *Proceedings of the 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST)*, pages 1–4, Thrissur, Kerala, India.

Received: October 21, 2024 Revised: December 17, 2024 Accepted: January 19, 2025

[3] J. A. Bartholomai and H. B. Frieboes. Lung cancer survival prediction via machine learning regression, classification, and statistical techniques. pages 632–637.

[4] Yasemin Gültepe. Performance of lung cancer prediction methods using different classification algorithms. *Computers, Materials & Continua*.

[5] Joy Chakra Bortty, Proshanta Kumar Bhowmik, Syed Ali Reza, Irin Akter Liza, Mohammed Nazmul Islam Miah, Muhammad Shoyaibur Rahman Chowdhury, and Md Al Amin. Optimizing lung cancer risk prediction with advanced machine learning algorithms and techniques. *Journal of Modern Health Sciences*.

[6] Qinglan Ma, Yulong Shen, Wei Guo, Kaiyan Feng, Tao Huang, and Yudong Cai. Machine learning reveals impacts of smoking on gene profiles of different cell types in lung. *Life*, 14(502).

[7] R. K. Pathan, I. J. Shorna, M. S. Hossain, M. U. Khandaker, H. I. Almohammed, and Z. Y. Hamd. The efficacy of machine learning models in lung cancer risk prediction with explainability. *PLoS ONE*, 19(6):e0305035.

[8] E. Nemlander, A. Rosenblad, E. Abedi, S. Ekman, J. Hasselstrom, and L. E. Eriksson. Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, former smokers and current smokers. *PLoS ONE*, 17(10):e0276703.

[9] J. C. Hsu, P.-A. Nguyen, P. T. Phuc, T.-C. Lo, M.-H. Hsu, M.-S. Hsieh, N. Q. K. Le, C.-T. Cheng, T. H. Chang, and C.-Y. Chen. Development and validation of novel deep-learning models using multiple data types for lung cancer survival. *Cancers*, 14(5562).

[10] E. Dritsas and M. Trigka. Lung cancer risk prediction with machine learning models. *Big Data Cogn. Comput.*, 6(139).

[11] K. Balachandran and R. Anitha. Ensemble based optimal classification model for pre diagnosis of lung cancer. In *IEEE*, page 31661.

[12] K. Kobylińska, T. Orłowski, M. Adamek, and P. Biecek. Explainable machine learning for lung cancer screening models. *Applied Sciences*, 12(1926).

[13] Huu-Huy Ngo and Hung Linh Le. A prediction model for lung cancer levels based on machine learning. *International Journal of Open Information Technologies*, 10(5):202.

[14] Saif Al Rumhi, Raza Hasan, Saqib Hussain, and Jitendra Pandey. Lung cancer prediction using machine learning techniques. In *6th Middle East College Student Research Conference Proceeding*.

[15] Radhanath Patra. Prediction of lung cancer using machine learning classifier. *Electronics Science, Berhampur University, Berhampur, Odisha, India*.

[16] Md Nur Hossain, Murshida Alam, Nafis Anjum, Md Habibur Rahman, Md Siam Taluckder, Md Nad Vi Al Bony, S M Shadul Islam Rishad, and Afrin Hoque Jui. Performance of machine learning algorithms for lung cancer prediction: a comparative study.

[17] Masoud Aliramezani, Charles Robert Koch, and Mahdi Shahbakhti. Modeling, diagnostics, optimization, and control of internal combustion engines via modern machine learning techniques: A review and future directions. *Progress in Energy and Combustion Science*, 88:100967, January 2022.

[18] Heng Chi, Yuyu Zhang, Tsz Ling Elaine Tang, Lucia Mirabella, Livio Dalloro, Le Song, and Glaucio H. Paulino. Universal machine learning for topology optimization. *Computer Methods in Applied Mechanics and Engineering*, 375:112739, March 2021.

[19] Kerr Ding, Michael Chin, Yunlong Zhao, Wei Huang, Binh Khanh Mai, Huanan Wang, Peng Liu, Yang Yang, and Yunan Luo. Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering. *Nature Communications*, 15(1):6392, July 2024. Publisher: Nature Publishing Group.

[20] Ponnarasan Krishnan. Ai-Driven Optimization In Healthcare: Machine Learning Models For Predictive Diagnostics And Personalized Treatment Strategies. *Well Testing Journal*, 33(S2):10–33, September 2024. Number: S2.

Received: October 21, 2024 Revised: December 17, 2024 Accepted: January 19, 2025

[21] Shiqi Wang, Peng Xia, Keyu Chen, Fuyuan Gong, Hailong Wang, Qinghe Wang, Yuxi Zhao, and Weiliang Jin. Prediction and optimization model of sustainable concrete properties using machine learning, deep learning and swarm intelligence: A review. *Journal of Building Engineering*, 80:108065, December 2023.

[22] Morteza Esfandyari, Amin Amiri Delouei, and Ali Jalai. Optimization of ultrasonic-excited double-pipe heat exchanger with machine learning and PSO. *International Communications in Heat and Mass Transfer*, 147:106985, October 2023.

[23] Binayak Kar, Widhi Yahya, Ying-Dar Lin, and Asad Ali. Offloading Using Traditional Optimization and Machine Learning in Federated Cloud–Edge–Fog Systems: A Survey. *IEEE Communications Surveys & Tutorials*, 25(2):1199–1226, 2023. Conference Name: IEEE Communications Surveys & Tutorials.

Received: October 21, 2024 Revised: December 17, 2024 Accepted: January 19, 2025