Applied Science and Biotechnology Journal for Advanced Research e-ISSN: 2583-553X
Peer Reviewed Journal
Volume-3, Issue-5 (September 2024), Page: 31-36 https://doi.org/10.5281/zenodo.14015999

LLM Machine Learning for Predicting Cardiovascular Mortality in Patients

Yue Zhu[1*], Xiaoyi Zhang[2] and Yuechen Zhang[3]
[1]Georgia Institute of Technology, USA
[2]Jocobi Medical Center, CHINA
[3]Mailman School of Public Health, Columbia University, USA

[*]**Corresponding Author:** Yue Zhu

Received: 20-08-2024 Revised: 05-08-2024 Accepted: 29-09-2024

ABSTRACT
Patients with chronic kidney disease (CKD) face a high risk of cardiovascular death, yet accurately predicting this risk remains challenging. This study aims to develop an interpretable machine learning (ML) model to predict 10-year cardiovascular mortality in CKD patients using SHAP explainers. [1]Six ML models were created and tested, with the best model selected for prediction and patient categorization. Survival rates were analyzed using log-rank tests on Kaplan-Meier curves, and Cox regression was employed to explore the relationship between ML-predicted risk scores and mortality. The chosen autoencoder (AE) model demonstrated superior performance, with higher ML scores[2] significantly correlating with increased cardiovascular mortality risk. Key determinants such as age, high blood pressure, C-reactive protein, and serum creatinine were identified. The ML-driven tool showcased high accuracy in determining the 10-year cardiovascular mortality risk for CKD patients, offering valuable insights for individual risk assessments.

Keywords: cardiovascular mortality, chronic kidney disease, machine learning, interpretable model, shap

I. INTRODUCTION

Machine learning (ML) has seen extensive adoption in the medical domain, revolutionizing healthcare technologies (12).[3] Within ML for the medical field, cardiovascular risk prediction remains a popular research area, where ML models have emerged as prominent tools, often surpassing traditional methods in risk assessment and stratification. In particular, recent studies report ML outperforming conventional methods in tasks such as atherosclerotic plaque tissue characterization[4,5,6,7,8,9], stratification of CVD risk [10,11,12,13,14,15], prediction of stroke risk, and forecasting of cardiac events (6). However, a significant gap persists in applying these advancements to understand cardiovascular implications in CKD patients. Given the continued increase in CKD prevalence and incidence it is essential to provide more focus on cardiovascular Specifically, this study integrated an advanced ML model with a framework based on SHAP. This approach not only enhances the accuracy of predicting 10-year cardiac mortality risk in CKD patients. Additionally, it assists clinicians in daily routine associated with the assessment of disease severity, thereby optimizing the potential for early intervention. Such developments represent a significant advancement for ML in medicine (16,17,20) and contribute to establishing interpretable and individualized models for risk prediction.

Related work

The global rise in CVD incidence and mortality remains a major obstacle to healthy ageing [19]. Current research prioritizes early detection, management, and treatment of CVD [20]. Advanced statistical and machine learning techniques have improved predictive health models [21]. More researchers utilize machine learning for cardiac mortality risk prediction, leveraging diverse models to assess mortality causes by CVD, from individual to composite risk factors [22]. While numerous studies have employed ML to predict and analyze cardiovascular mortality, reports on CVD within the context of CKD remain hardly explored. It is noteworthy that CKD induces a systemic, chronic proinflammatory state [23]. This state plays a pivotal role in vascular and myocardial remodeling, contributing to accelerated atherosclerotic lesion development, vascular calcification and senescence, as well as myocardial fibrosis and cardiac valve calcification. Consequently, CKD can be perceived as causing an expedited aging of the cardiovascular system. Given this, research on risk prediction of cardiac death within CKD is of paramount importance [16]. To this end, several common ML algorithms include Naive Bayesian, Support Vector Machine, KNN, and Multilayer perceptron etc. in our report. Yet, the "curse of dimensionality" in biological CVD data remains. Dimensionality reduction methods like Principal Component Analysis and Autoencoders[24] have become essential. In practical terms, achieving a balance between model precision and clarity is challenging, and making model outcomes more intuitive for medical professionals is necessary [25].

II. METHODS

2.1 Study Population

The study analyzed the data from the combination of five continuous survey circles of the National Health and Nutrition Examination Survey (NHANES), spanning from 2001 to 2010. The inclusion criteria for the current research were as follows: (1) age ≥20 years; (2) diagnosed with CKD (defined as an estimated glomerular filtration rate [eGFR] <60 ml/min/1.73 m2[using the Chronic Kidney Disease-Epidemiology Collaboration (CKD-EPI) equation (26) and/or a urinary albumin-Cr ratio [ACR] >30 mg/g); (3) possessing complete baseline, and follow-up data. Ultimately, date from a cohort of 2,935 individuals were used in the present research (Figure 1).

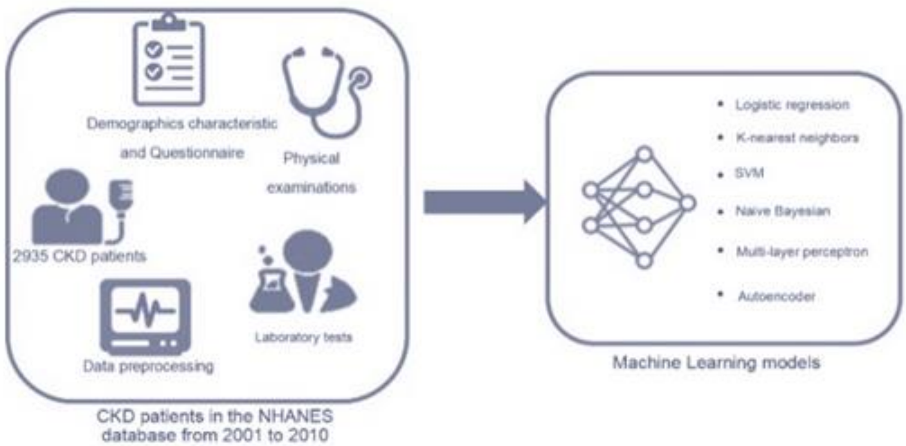
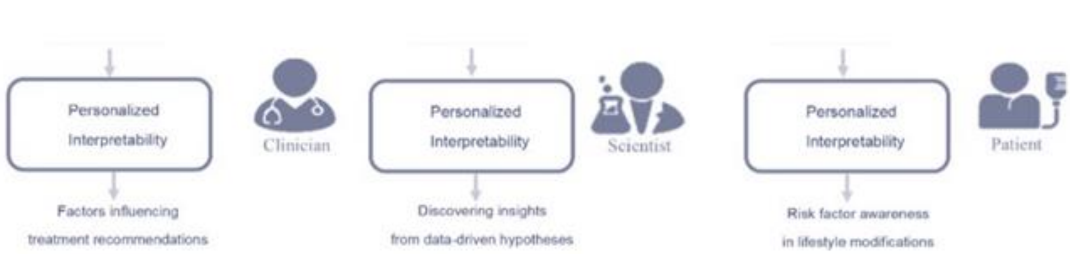


Figure 1: Framework for predicting cardiovascular mortality in patients with chronic kidney disease. SVM: support vector machines, SHAP: SHapley Additive exPlanations.

2.2 Data Source

All data utilized in this research were sourced from NHANES, an ongoing cross-sectional survey designed to capture nationally representative samples of the non-institutionalized population in the United States. Data were collected through the official website of the Centers for Disease Control and Prevention [https://wwwn.cdc.gov/Nchs/Nhanes]. The NHANES sampling frame is established on a complex, stratified, and multi-stage probability sample design. Data were extracted from various files, each encompassing a specific set of variables for a given year. Integration of these files was performed to construct a consolidated database containing the entirety of available data for each individual (with the SEQN identifier serving as the linkage across all datasets). This approach facilitated the generation of a unique and comprehensive database encompassing the entire cohort of examined subjects, along with their related data. The Ethics Review Board of the National Center for Health Statistics granted approval for all NHANES protocols, and all participants provided written informed consent.



2.3 Feature Selection

Our structured database encompassed 33 variables, often referred to as “features” in ML, which were selected to be associated with the cause or progression of CKD based on domain knowledge and literature search. Features exhibiting a missing data rate not exceeding 30% were retained and subjected to multivariate feature imputation to in the final

Sensitivity	0.7444	0.7218	0.5639	0.4812	0.7594	0.8095
	(0.7091,0.7797)	(0.6856,0.7581)	(0.5238,0.6040)	(0.4408,0.5216)	(0.7248,0.7940)	(0.7765,0.8425)
Specificity	0.9758	0.9824	0.9846	0.9868	0.9251	0.9634
	(0.9633,0.9882)	(0.9717,0.9930)	(0.9746,0.9945)	(0.9775,0.9960)	(0.9038,0.9464)	(0.9357,0.9912)
F1-score	0.8148	0.8101	0.6977	0.6305	0.7537	0.7000
	(0.7834,0.8462)	(0.7784,0.8419)	(0.6605,0.7348)	(0.5915,0.6696)	(0.7189,0.7886)	(0.6629,0.7371)
	0.0627	0.0628	0.0930	0.0949	0.0909	0.1206
	(0.0431,0.0823)	(0.0432,0.0825)	(0.0695,0.1165)	(0.0712,0.1186)	(0.0676,0.1141)	(0.0943,0.1469)

2.4 Model Development

Six ML models were formulated to predict 10-year cardiovascular mortality leveraging follow-up data. In addition to the five common models sourced from Scikit-learn (28), namely, logistic regression (LR), k-nearest neighbors (KNN), multi-layer perceptron (MLP), support vector machines (SVM), Naive Bayesian (NB), and k-nearest neighbors (KNN), we incorporated autoencoders (AE) [29]. AE emerges as an innovative learning technique capable of ensuring data-driven dimensionality reduction with minimal prior the endpoint of cardiac mortality we employed stratified random sampling to partition 2,935 patients into training and test sets, maintaining a 7: 3 ratio. The training dataset was preprocessed via the Adaptive Synthetic Sampling Approach (ADASYN) (30), assuring a balanced distribution between the minority and majority classes. The optimization of ML model parameters was achieved through ten-fold cross-validation.

2.5 Model Interpretation for Healthcare

In healthcare, the interpretability of machine learning models influences clinical decision-making. Understanding prediction drivers is important for patient outcomes, but machine learning models, dubbed 'black boxes' can cloud their logic in fields like healthcare[30] To bridge this chasm, we introduced the SHAP values into our research, presented by Lundberg and Lee (31), SHAP offers a consolidated structure tailored to elucidate the predictions of ML models, presenting an easy to grasp approach to various intricate ML algorithms in an e.g. routine clinical setting. Its prowess in enhancing interpretability, a pivotal asset in healthcare, has been corroborated in prior research [32-35]. Notably, SHAP facilitates both fine- and coarse-grained model interpretability, setting it apart with a more theoretical foundation than many of its previous published the endpoint of cardiac mortality we employed stratified random sampling to partition 2,935 patients into training and test sets, maintaining a 7: 3 ratio. The training dataset was preprocessed via the Adaptive Synthetic Sampling Approach (ADASYN) (30), assuring a balanced distribution between the minority and majority classes. The optimization of ML model parameters was achieved through ten-fold cross-validation.

2.6 Model Interpretation for Healthcare

In healthcare, the interpretability of machine learning models influences clinical decision-making. Understanding prediction drivers is important for patient outcomes, but machine learning models, dubbed 'black boxes' can cloud their logic in fields like healthcare (30). To bridge this chasm, we introduced the SHAP values into our research, presented by Lundberg and Lee (31), SHAP offers a consolidated structure tailored to elucidate the predictions of ML models, presenting an easy to grasp approach to various intricate ML algorithms in an e.g. routine clinical setting. Its prowess in enhancing interpretability, a pivotal asset in healthcare, has been corroborated in prior research (32-35). Notably, SHAP facilitates both fine- and coarse-grained model interpretability, setting it apart with a more theoretical foundation than many of its previous published[60,61,62,63,64,65]

2.7 Experimental Implementation and Statistical Analysis

Analyses were conducted utilizing Python (Version 3.9) with the integration of several libraries including Imblearn, Sklearn, Matplotlib, Lifelines, Shap, and Tensorflow. Additionally, R (Version 4.2.1) was employed, specifically leveraging the survival and survminer packages. To fully assess the capability of the ML models, various evaluation metrics were adopted, as described in [37], encompassing specificity, sensitivity, F1-score, and the area under the receiver operating characteristic curve (AUC). For comparative analysis, the evaluation metrics across the six models by one-way ANOVA, followed by post-hoc pair wise comparisons using the least-significant difference method. The optimal cut-off, as determined by maximizing the Youden's index, enabled the stratification of patients into low and high ML risk groups. Differences in survival outcomes were assessed via the log-rank test on the Kaplan–Meier curves. Furthermore, multivariable Cox regression was employed to evaluate the relationship between ML risk and 10-year cardiovascular death. All statistical inferences were made with a threshold for significance set at a two-tailed P value ≤ 0.05.[45,46,47,48,49,50]

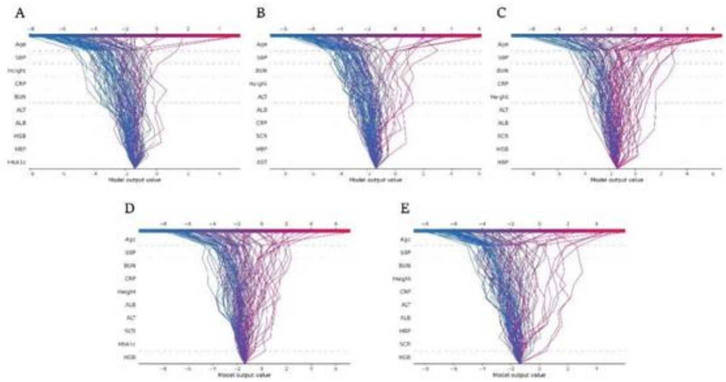


Figure 3: Variable importance in ML classification for race. (A) Mexican American, (B) of these influences resulted in the final SHAP value, corresponding to the prediction score.

III. CONCLUSION

In summary, the tested ML models have demonstrated satisfying performance in predicting 10-year cardiovascular mortality among patients with CKD. Our ML model effectively incorporates important clinical and demographic features to accurately discern cardiovascular mortality in CKD stage 3-5 population, thus facilitating personalized preventive strategies in clinic practice. The combination of ML and SHAP holds promise in delivering precise and transparent individualized risk predictions. This integration helps doctors understand key features in the model, improving their insight into the decision-making related to disease severity assessment.

REFERENCES

1. Shimizu, Shosei et al. (2021). Proton beam therapy for a giant hepatic hemangioma: A case report and literature review. *Clinical and Translational Radiation Oncology, 27*, 152-156. doi:10.1016/j.ctro.2021.01.014.
2. Shimizu, Shosei et al. (2023). Boron neutron capture therapy for recurrent glioblastoma multiforme: Imaging evaluation of a case with long-term local control and survival. *Cureus, 15*(1), e33898. doi:10.7759/cureus.33898.
3. Li, Yinuo et al. (2023). A retrospective study of renal growth changes after proton beam therapy for pediatric malignant tumor. *Current Oncology (Toronto, Ont.), 30*(2), 1560-1570. doi:10.3390/curroncol30020120.
4. Nakamura, Masatoshi et al. (2024). A systematic review and meta-analysis of radiotherapy and particle beam therapy for skull base chondrosarcoma: TRP-chondrosarcoma 2024. *Frontiers in Oncology, 14,* 1380716. doi:10.3389/fonc.2024.1380716.
5. Nitta, Hazuki et al. (2024). An analysis of muscle growth after proton beam therapy for pediatric cancer. *Journal of Radiation Research, 65*(2), 251-255. doi:10.1093/jrr/rrad105.
6. Jin, Yonglong et al. (2023). Proton therapy (PT) combined with concurrent chemotherapy for locally advanced non-small cell lung cancer with negative driver genes. *Radiation Oncology (London, England), 18*(1), 189. doi:10.1186/s13014-023-02372-8.
7. Li, Yinuo et al. (2023). Smart nanofiber mesh with locally sustained drug release enabled synergistic combination therapy for glioblastoma. *Nanomaterials (Basel, Switzerland), 13*(3), 414. doi:10.3390/nano13030414.
8. Saito, Takashi et al. (2024). Systematic review and meta-analysis of particle beam therapy versus photon radiotherapy for skull base chordoma: TRP-chordoma 2024. *Cancers, 16*(14), 2569. doi:10.3390/cancers16142569.
9. Li, Yinuo et al. (2024). Late changes in renal volume and function after proton beam therapy in pediatric and adult patients: Children show significant renal atrophy but deterioration of renal function is minimal in the long-term in both groups. *Cancers, 16*(9), 1634. doi:10.3390/cancers16091634.
10. Niitsu, Hikaru et al. (2024). Tumor response on diagnostic imaging after proton beam therapy for hepatocellular carcinoma. *Cancers, 16*(2), 357. doi:10.3390/cancers16020357.
11. Li, Yinuo et al. (2022). Proton beam therapy for multifocal hepatocellular carcinoma (HCC) showing complete response in pathological anatomy after liver transplantation. *Cureus, 14*(6), e25744. doi:10.7759/cureus.25744.
12. Kumada, Hiroaki et al. (2022). *Current development status of iBNCT001, demonstrator of a LINAC-based neutron source for BNCT*, 347–358. doi:10.3233/JNR-220029.
13. Wang, Randi, & Morad Behandish. (2022). *Surrogate modeling for physical systems with preserved properties and*

adjustable tradeoffs. arXiv preprint arXiv:2202.01139.

14. Wang, Randi, & Vadim Shapiro. (2019). Topological semantics for lumped parameter systems modeling. *Advanced Engineering Informatics, 42,* 100958.
15. Wang, Randi, Vadim Shapiro, & Morad Mehandish. (2024). Model consistency for mechanical design: Bridging lumped and distributed parameter models with a priori guarantees. *Journal of Mechanical Design, 146*(5).
16. Chen, M., Chen, Y., & Zhang, Q. (2021). A review of energy consumption in the acquisition of bio-feedstock for microalgae biofuel production. *Sustainability, 13*(16), 8873.
17. Chen, M. (2023). *Investigating the influence of interannual precipitation variability on terrestrial ecosystem productivity*. Doctoral Dissertation, Massachusetts Institute of Technology.
18. Chen, M. (2021, December). Annual precipitation forecast of Guangzhou based on genetic algorithm and backpropagation neural network (GA-BP). in *International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2021), 12156*, pp. 182-186. SPIE.
19. Dong, S., Xu, T., & Chen, M. (2022, October). Solar radiation characteristics in Shanghai. in *Journal of Physics: Conference Series, 2351*(1), pp. 012016. IOP Publishing.
20. Zhang, X., Soe, A. N., Dong, S., Chen, M., Wu, M., & Htwe, T. (2024). Urban resilience through green roofing: A literature review on dual environmental benefits. in *E3S Web of Conferences, 536*, pp. 01023. EDP Sciences.
21. Chen, M., Chen, Y., & Zhang, Q. (2024). Assessing global carbon sequestration and bioenergy potential from microalgae cultivation on marginal lands leveraging machine learning. *Science of the Total Environment, 948,* 174462.
22. Song, Y., Arora, P., Varadharajan, S. T., Singh, R., Haynes, M., & Starner, T. (2024, April). Looking from a different angle: Placing head-worn displays near the nose. in *Proceedings of the Augmented Humans International Conference*, pp. 28-45.
23. Song, Y., Arora, P., Singh, R., Varadharajan, S. T., Haynes, M., & Starner, T. (2023, October). Going blank comfortably: Positioning monocular head-worn displays when they are inactive. in *Proceedings of the 2023 ACM International Symposium on Wearable Computers*, pp. 114-118.
24. Qian, C., Guo, Y., Mo, Y., & Li, W. (2024). *WeatherDG: LLM-assisted procedural weather generation for domain-generalized semantic segmentation*. arXiv [Cs.CV]. Retrieved from: http://arxiv.org/abs/2410.12075.
25. Wang, Yang, Yojiro Mori, & Hiroshi Hasegawa. (2021). Dynamic routing and spectrum allocation based on actor-critic learning for multi-fiber elastic optical networks. *Photonics in Switching and Computing 2021, Optica Publishing Group*, pp. W1B.3. doi:10.1364/PSC.2021.W1B.3.
26. Wang, Yang, et al. (2018). Optimizing multi-criteria k-shortest paths in graph by a natural routing genotype-based genetic algorithm. *13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 341-345. doi:10.1109/ICIEA.2018.8397739.
27. Wang, Yang, Yojiro Mori, & Hiroshi Hasegawa. (2020). Resource assignment based on core-state value evaluation to handle crosstalk and spectrum fragments in sdm elastic optical networks. *Opto-Electronics and Communications Conference (OECC),* pp. 1-3. doi:10.1109/OECC48412.2020.9273621.
28. Zhao, Yuwen, Baojun Hu, & Sizhe Wang. (2024). *Prediction of brent crude oil price based on LSTM model under the background of low-carbon transition*. arXiv preprint arXiv:2409.12376.
29. Yan, Hao, et al. (2024). *Research on image generation optimization based deep learning*.
30. Tang, Xirui, et al. (2024). *Research on heterogeneous computation resource allocation based on data-driven method*. arXiv preprint arXiv:2408.05671.
31. Han, Yi, & Thomas CM Lee. (2022). Uncertainty quantification for sparse estimation of spectral lines. *IEEE Transactions on Signal Processing, 70*, 6243-6256.
32. Yao, Jiawei, et al. (2023). Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society.
33. Yao, Jiawei, et al. (2024). *QE-BEV: Query evolution for bird's eye view object detection in varied contexts*. ACM Multimedia.
34. Pan, Xiaochao, et al. (2024). *HarmonicNeRF: Geometry-informed synthetic view augmentation for 3d scene reconstruction in driving scenarios*. ACM Multimedia.
35. Ma, B., Ma, B., Gao, M., Wang, Z., Ban, X., Huang, H., & Wu, W. (2021). Deep learning based automatic inpainting for material microscopic images. *Journal of Microscopy, 281*(3), 177-189.
36. Wang, Y., Ban, X., Wang, H., Li, X., Wang, Z., Wu, D., ... & Liu, S. (2019). Particle filter vehicles tracking by fusing multiple features. *IEEE Access, 7*, 133694-133706.
37. Liu, Jiabei, et al. (2024). Application of deep learning-based natural language processing in multilingual sentiment analysis. *Mediterranean Journal of Basic and Applied Sciences (MJBAS), 8*(2), 243-260.
38. Xu, Qiming, et al. (2024). Applications of explainable ai in natural language processing. *Global Academic Frontiers,*

2(3), 51-64.

39. Zhong, Yihao, et al. (2024). Deep learning solutions for pneumonia detection: Performance comparison of custom and transfer learning models. *medRxiv*, 2024-06.
40. Zhu, Armando, et al. (2024). *Exploiting diffusion prior for out-of-distribution detection*. arXiv preprint arXiv:2406.11105.
41. Li, Keqin, et al. (2024). *Exploring the impact of quantum computing on machine learning performance*.
42. Gu, Wenjun, et al. (2024). *Predicting stock prices with FinBERT-LSTM: Integrating news sentiment analysis*. arXiv preprint arXiv:2407.16150.
43. Wang, Zixiang, et al. (2024). *Research on autonomous driving decision-making strategies based deep reinforcement learning*. arXiv preprint arXiv:2408.03084.
44. Bo, Shi, et al. (2024). *Attention mechanism and context modeling system for text mining machine translation*. arXiv preprint arXiv:2408.04216.
45. Qian, Yang, et al. (2020). Heterogeneous optoelectronic characteristics of Si micropillar arrays fabricated by metal-assisted chemical etching. *Scientific Reports*, *10*(1), 16349.
46. Li, Wei, et al. (2018). An intelligent electronic lock for remote-control system based on the internet of things. *Journal of Physics: Conference Series*, *1069*(1). IOP Publishing.
47. Gao, Haoqi, et al. (2016). A novel texture extraction method for the sedimentary structures' classification of petroleum imaging logging. *Pattern Recognition: 7th Chinese Conference, Chengdu, China, Proceedings, Part II 7*. Springer Singapore.
48. Yan, Hao, et al. (2024). *Research on image generation optimization based deep learning*.
49. Tang, Xirui, et al. (2024). *Research on heterogeneous computation resource allocation based on data-driven method*. arXiv preprint arXiv:2408.05671.
50. Su, Pei-Chiang, et al. (2022). A mixed-heuristic quantum-inspired simplified swarm optimization algorithm for scheduling of real-time tasks in the multiprocessor system. *Applied Soft Computing, 131*, 109807.
51. Zhao, Yuwen, Baojun Hu, & Sizhe Wang. (2024). *Prediction of brent crude oil price based on LSTM model under the background of low-carbon transition*. arXiv preprint arXiv:2409.12376.
52. Diao, Su, et al. (2024). *Ventilator pressure prediction using recurrent neural network*. arXiv preprint arXiv:2410.06552.
53. Zhao, Qinghe, Yue Hao, & Xuechen Li. (2024). *Stock price prediction based on hybrid CNN-LSTM model*.
54. Yin, Ziqing, Baojun Hu, & Shuhan Chen. (2024). *Predicting employee turnover in the financial company: A comparative study of CatBoost and XGBoost Models*.
55. Diao, Su, et al. (2024). *Ventilator pressure prediction using recurrent neural network*. arXiv preprint arXiv:2410.06552.
56. Qian, Chenghao, et al. (2024). *WeatherDG: LLM-assisted procedural weather generation for domain-generalized semantic segmentation*. arXiv preprint arXiv:2410.12075.
57. Li, Zhenglin, et al. (2023). Stock market analysis and prediction using LSTM: A case study on technology stocks. *Innovations in Applied Engineering and Technology*, 1-6.
58. Mo, Yuhong, et al. (2024). Large Language Model (LLM) AI text generation detection based on transformer deep learning algorithm. *International Journal of Engineering and Management Research, 14*(2), 154-159.
59. Li, Shaojie, Yuhong Mo, & Zhenglin Li. (2022). Automated pneumonia detection in chest x-ray images using deep learning model. *Innovations in Applied Engineering and Technology*, 1-6.
60. Mo, Yuhong, et al. (2024). Password complexity prediction based on roberta algorithm. *Applied Science and Engineering Journal for Advanced Research, 3*(3), 1-5.
61. Liu, Tianrui, et al. (2024). Spam detection and classification based on distilbert deep learning algorithm. *Applied Science and Engineering Journal for Advanced Research, 3*(3), 6-10.
62. Dai, Shuying, et al. (2024). The cloud-based design of unmanned constant temperature food delivery trolley in the context of artificial intelligence. *Journal of Computer Technology and Applied Mathematics, 1*(1), 6-12.
63. Mo, Yuhong, et al. (2024). Make scale invariant feature transform "Fly" with CUDA. *International Journal of Engineering and Management Research, 14*(3), 38-45.
64. He, Shuyao, et al. (2024). Lidar and monocular sensor fusion depth estimation. *Applied Science and Engineering Journal for Advanced Research, 3*(3), 20-26.
65. Liu, Jihang, et al. (2024). Unraveling large language models: From evolution to ethical implications-introduction to large language models. *World Scientific Research Journal, 10*(5), 97-102.
66. Mo Yuhong, Zhang Yuchen, Li Hanzhe, Wang Han, & Yan Xu. (2024). Prediction of heart failure patients based on multiple machine learning algorithms. *Applied and Computational Engineering, 75*. 1-7. doi:10.54254/2755-2721/75/20240498.