Data-Driven Model for Optimizing Active Learning

Abang Asyraaf Aiman Abang Azahari, Marshima Mohd Rosli, Nor Shahida Mohamad Yusop
Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA, 40450, Selangor, Malaysia

***Abstract*—Data-driven models depend on extensive datasets for precise predictions; yet, acquiring adequate labeled data for training these models is a challenge, especially with medical datasets that are constrained by privacy considerations, resulting in a deficiency of labeled data. Active Learning (AL) has developed as a cost-effective strategy that minimizes the quantity of labeled data required for training by selecting the most informative samples. The performance of active learning methods is significantly influenced by data quality characteristics, and due to a lack of direction in selecting the most suitable active learning approach. The study presents a data-driven selection approach that suggests appropriate active learning methods based on dataset characteristics. The study examines the characteristics of the dataset and their impact on active learning performance, revealing significant correlations between data quality issues and the efficacy of active learning approaches. A rule-based selection model is subsequently constructed and verified by experiments and case studies across various datasets. The findings demonstrated consistent alignment between suggested and practically effective techniques. Statistical analysis verifies that the data-driven selection model exhibits reliability exceeding chance agreement, indicating its robustness and practical application in recommending AL techniques selection.**

***Keywords*—Data-driven models; selection model; data quality characteristics; active learning**

I. Introduction

Data-driven models are widely used in various domains, such as finance, healthcare, and marketing, to make data-driven decisions. These models often require large volumes of labeled data and statistical methods to determine the most appropriate machine learning methods for a given data set and prediction problem [1]. These models embedded optimization techniques to enable quick forecasting by using relevant information during model training, contributing to their popularity in predictive tasks [2]. This approach can help overcome the limitations of traditional model selection methods, which rely on predetermined formulas rather than empirical data [3], [4].

Data-driven models learn from actual datasets, identifying patterns and relationships inherent in the data. However, the performance of data-driven models is highly dependent on data quality, as they require large volumes of high-quality labeled data. Issues such as limited data availability, biased samples, and high dimensionality can significantly impair model effectiveness. These challenges are critical as inconsistent data can severely affect a model's generalizability and predictive accuracy [5], [6].

To mitigate these drawbacks, Active Learning (AL) has emerged as a promising solution that reduces the need for extensive labeled datasets by iteratively selecting the most informative samples for annotation, thereby enhancing model performance with fewer labeled instances. This method is valuable in domains such as medical research, where accurate data selection is crucial and labeling is costly [6]. AL is also vulnerable to data quality issues, including excessive dimensionality, sample bias, and insufficient data, which can diminish its effectiveness [7], [8].

Selecting appropriate AL methods is important in building an effective predictive model because it can significantly improve the efficiency and accuracy of the learning process. AL involves iteratively selecting the most informative samples from a large unlabeled dataset to be labeled by an expert and using these labeled samples to train a predictive model [9], [10]. There are many different types of active learning methods, each with its own strengths and weaknesses, and the choice of method can have a significant impact on the accuracy and generalizability of the resulting predictive model [11].

However, there is currently no validated approach that can recommend the best method for any type of dataset, as the optimal choice of active learning depends on several factors, including the size and complexity of the dataset, the characteristics of the input and output variables, and the specific domain of the prediction model. In addition, the performance of a predictive model may be influenced by other factors, such as the quality of the data, the choice of hyperparameters, and the specific implementation of the AL method [12], [13], [14].

Building a predictive model with AL methods without analyzing the domain and characteristics of the dataset may result in lower prediction performance [15]. This problem occurs because some active learning methods may not select the most informative or representative data for labeling if the data preprocessing techniques used do not effectively represent the data or do not consider the unique features and patterns of the data [16][10]. For example, if the data preprocessing techniques used do not effectively represent the data, the active learning method may not be able to select the most informative samples for labeling.

Previous studies have reported that not all AL methods are able to make accurate predictions for different characteristics of datasets [8], [17], [18]. Research has shown that experimental studies that use active learning methods without analyzing the characteristics of the dataset and the domain of the dataset result in unpredictable performance of prediction model [8], [19], [20]. While some of the commonly used learning methods have been proposed, one problem that has not received the attention it deserves is determining appropriate AL methods for a given dataset and domain characteristics. This motivates the development of a data-driven selection model that recommends active learning methods based on dataset characteristics.

The study is structured as follows: Section II reviews related work, Section III describes the methodology, Section IV presents the results and discussion, and Section V concludes with key findings and future directions.

II. Related Work

Adoption of data-driven selection models is increasing due to their effectiveness in producing accurate and efficient machine learning predictions. The optimization techniques embedded within these models enable quick forecasting by using relevant information during training, which contributes to their popularity in predictive tasks [2]. This contrasts with model-driven approaches, which rely on predetermined formulas or abstract theoretical constructs rather than empirical data. Data-driven models learn directly from actual data, identifying patterns such as regularities, trends, or structures and relationships inherent in the dataset. This fundamental difference accounts for the observed variations in both computational efficiency and predictive accuracy between the two approaches [21].

Bernhardt et al. propose a data-driven strategy, called "active label cleaning", to prioritize samples for re-annotation [22]. The proposed approach involves ranking instances based on estimated label correctness and labeling difficulty, and a simulation framework is introduced to assess relabeling efficacy. The experiments conducted on natural images and a medical imaging benchmark created specifically for the research demonstrate that cleaning noisy labels can reduce their negative impact on model training, evaluation, and selection.

Mahapatra et al. applied Interpretability-Driven Sample Selection (IDEAL), a self-supervised learning-based approach to train a classifier that identifies the most informative sample in a given batch of images [23]. IDEAL is demonstrated in an active learning setup for lung disease classification and histopathology image segmentation. The results show that the proposed self-supervised approach outperforms other methods for selecting informative samples, leading to state-of-the-art performance with fewer samples.

In healthcare practices, the outcome of a data-driven prediction model highly relies on the data that it learns. However, if the quality of the medical data that it learns is imbalanced, then the biased model can make such a classification error [24]. Hence, diagnosing a medical condition can create a false alarm result where patients are incorrectly diagnosed, but the patient actually has a negative result, or worse, the other way around. This is undoubtedly true in real-life situations where the outcome of medical diagnostics determines patients' lives and deaths [25].

For building a disease prediction model, choosing a reliable active learning method is one of the critical challenges due to various factors influencing the selection process, particularly the variability characteristics of the dataset. Previous studies have assessed various active learning methods for disease prediction and discovered that for different types of datasets, the performance of the prediction methods also varies [17], [26]. These studies considered outliers and noise characteristics of the medical datasets and investigated the performance of the active learning methods [24].

Many studies have compared different active learning methods for prediction problems and found that characteristics of the dataset have a strong impact on the performance of the prediction model in different domains. The active learning method that performs best for a particular dataset may perform poorly for another dataset [27], [28]. Similarly, in healthcare, many studies have evaluated different active learning methods for disease prediction and found that prediction performance also varies for different types of datasets. This suggests that there is no single best method that can be used with any type of dataset for disease prediction [12].

One possible reason for the unpredictable performance of the prediction is that most studies have used active learning as a black box without analyzing the domain and the characteristics of the dataset. The prediction model can perform best if appropriate learning methods are selected for the right characteristics of the dataset. The research gap is that it is not clear to what extent the characteristics of the dataset affect the performance of the learning methods and how to select the appropriate method for an improved prediction process [12], [24]. This raises the need for a data-driven selection model that can suggest active learning methods for building an effective prediction model based on the characteristics of the dataset.

III. Methodology

This study has constructed a workflow for developing a data-driven selection model, as illustrated in Fig. 1. The workflow of model development has three main phases: Elements Identification, Model Construction, and Model Evaluation.

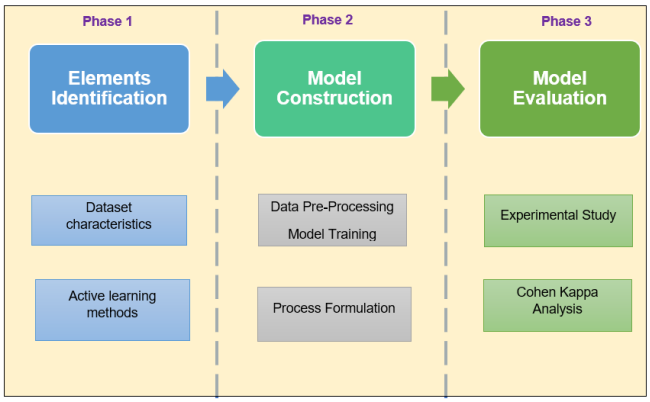

Fig. 1. Workflow of model development.

Phase 1 focuses on determining the dataset characteristics and active learning methods from the literature and datasets collection. The study performed a systematic mapping study to elicit the various characteristics of the dataset and the appropriateness of the active learning algorithms for the identified dataset characteristics from the existing studies. The study also determines the most potential active learning methods used in existing studies for building a prediction model. This phase establishes the essential elements required to formulate a model that aligns dataset properties with suitable active learning methods.

Phase 2 aims to construct a data-driven selection model for recommending active learning methods to build prediction model. This phase identified eight datasets that are commonly

used with five selected active learning methods. This phase starts with dataset pre-processing, which includes data cleaning, normalization, and feature selection to ensure data quality. Then, the study conducted model training using selected active learning methods with selected datasets. To ensure proper evaluation, the datasets were separated into independent training and testing stages: active learning models were trained on the labeled subset generated during AL iterations, while the remaining unseen data were reserved for testing. In the process formulation, the identified elements in Phase 1 are tabulated in a matrix table to show the association and relationship between elements. The process formulation then defines the operational process based on a recommender approach that links dataset characteristics with active learning methods, forming a coherent mechanism for developing a data-driven selection model.

Phase 3 focuses on the evaluation of the effectiveness of a data-driven selection model using an experimental study. This experimental study aims to establish the suitability of the different active learning methods for the different values of the datasets characteristics. The study also utilizes case studies to establish the effectiveness of the proposed model by comparing the recommendation of the proposed active learning methods from the proposed model with the results from the corresponding experimental studies. Based on experimental studies, the Cohen's Kappa analysis is used to measure inter-rater agreement between case studies and experimental results. While alternative metrics like accuracy or F1-score are useful for evaluating classification performance, they do not account for chance agreement. The high kappa values would indicate strong agreement between the case studies and experimental results in recommending active learning methods, and this would suggest the reliability of the model.

IV. Results and Discussion

This study aims to construct a data-driven selection model to facilitate the selection of the most reliable active learning methods in developing prediction models that are able to provide practitioners with an accurate decision support tool. The study determines the dataset characteristics that correlate with the active learning methods to build a prediction model. The study also evaluates the effectiveness of the proposed data-driven selection model using Cohen's Kappa analysis to ensure the agreement between case studies and experimental results.

A. Data-Driven Selection Model

The development of the data-driven selection model employs a recommendation approach to automate the identification of suitable active learning strategies for a certain dataset. Fig. 2 shows the conceptual framework for the data-driven selection model. The framework consists of two major components: Dataset Analysis and Active Learning Strategy Recommender. In the Dataset Analysis stage, data quality issues such as missing data, imbalance, high dimensionality, redundancy, and outliers are identified. These identified issues are then passed as inputs to the Data Driven Selection Model, which maps specific data quality characteristics to appropriate active learning strategies. In the Active Learning Strategy Recommender stage, the most suitable active learning methods are suggested based on the mapping results. This conceptual framework demonstrates the logical flow from dataset quality assessment to strategy recommendation, emphasizing a systematic and data-driven approach for optimizing active learning performance.

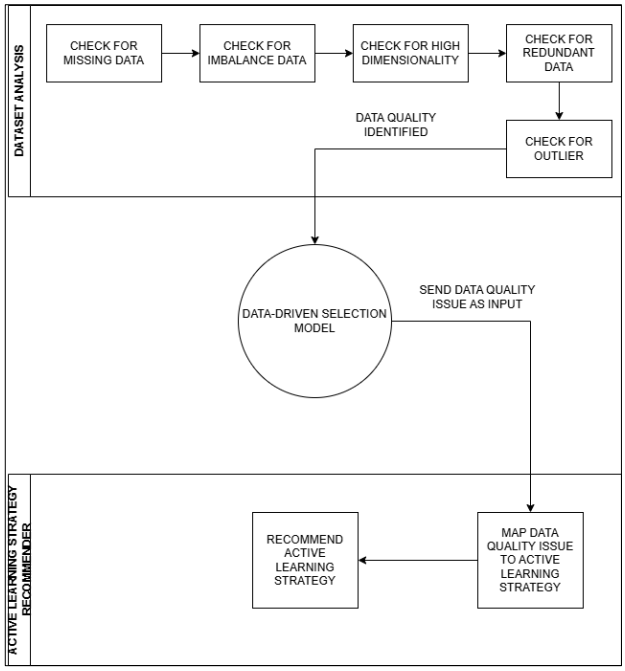

Fig. 2. Conceptual framework of the data-driven active learning strategy selection model.

B. Prototype of Data-Driven Selection Model

This study developed a prototype, Active Learning Strategy Recommender, to demonstrate the practical implementation of the proposed data-driven selection model. The prototype serves as a decision-support tool that recommends suitable active learning strategies based on data quality issues detected within a given dataset. It operates the conceptual framework by integrating data quality analysis, rule-based mapping, and strategy recommendations into a prototype. Fig. 3 shows the main user interface for the prototype.

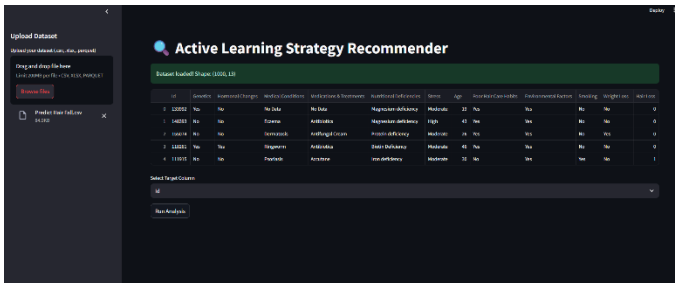

Fig. 3. Main user interface of the prototype.

As shown in Fig. 3, the main user interface allows users to begin by uploading a dataset file. Once the dataset is uploaded, the prototype conducts a data quality assessment to identify potential issues such as missing data, class imbalance, high dimensionality, redundancy and outliers. These assessments are performed according to the data quality characteristics. Users can initiate the analysis process by clicking the Run Analysis button, prompting the prototype to analyses the dataset and display the detected quality issues.

The prototype then applies the data-driven selection model to map the detected data quality issues to the most suitable active learning strategies. The recommended strategies' results are shown in Table I, displaying the recommended strategy consistent with the dataset's characteristics. The prototype also displays the alternative strategies that are listed according to their relevance based on the findings of this study.

TABLE I. RECOMMENDED ACTIVE LEARNING STRATEGY

Strategy	Score	Reason
UNC	1	Missing Value
QBC	0.993	Missing Value
DIV	0.949	Missing Value

C. Model Training

The performance results of the active learning strategies across medical datasets are summarized in Table II. Each dataset was evaluated under five data quality issues: high dimensionality (HD), imbalance, missing values, outliers, and redundancy. The active learning strategies compared are Uncertainty Sampling (UNC), Query by Committee (QBC), Diversity Sampling (DV), Density Sampling (DS), and Cluster-based Sampling (CS). The results show the classification accuracy achieved by each strategy for the respective data quality issues.

Table II illustrates the performance of active learning strategies applied to selected data quality issues. The active learning strategies that are Uncertainty sampling (UNC) and query-by-committee (QBC) indicate consistent performance (typically >0.90) for high-dimensional (HD), imbalanced, and redundant data. This shows that these strategies are reliable active learning methods for the common medical data challenges. However, both strategies show significant limitations when handling outliers, with performance often dropping below 0.50. Diversity-based sampling performs comparably to UNC and QBC in most scenarios while showing slightly better adaptability to missing data.

Cluster and density-based strategies were excluded from further consideration due to their inconsistent and often poor performance across multiple data characteristics. While these approaches occasionally showed adequate results in specific scenarios, as shown in the results where density-based sampling achieved 0.96 for the asthma dataset in the high dimensionality variant. Their overall performance was distinctly inferior to UNC, QBC, and diversity-based methods. From the results in Table I, it can be seen that cluster and density methods often obtained below 0.30, indicating these methods are unreliable for practical applications. This performance gap justifies focusing on the more robust UNC, QBC, and diversity-based strategies.

These findings suggest several practical recommendations: 1) UNC and QBC should be prioritized for HD, imbalance, or redundant medical data, given their consistently high accuracy. 2) Diversity-based sampling presents a viable alternative, particularly for datasets with missing values, where it sometimes outperforms other methods. 3) Future research should explore hybrid methods that combine the strengths of these top-performing strategies, especially to address challenging cases like outliers, where current methods still underperform.

TABLE II. MODEL TRAINING RESULTS

Dataset	Dataset quality issues	UNC	QBC	DV	DS	CS
Alzheimer	HD	0.90	0.97	1.00	0.67	0.63
	Imbalance	0.94	0.98	0.95	0.66	0.61
	Missing	0.86	0.91	0.95	0.69	0.62
	Outlier	0.44	0.75	0.47	0.30	0.26
	Redundancy	0.93	0.96	0.97	0.69	0.63
Asthma	HD	1.00	1.00	1.00	0.96	0.95
	Imbalance	0.95	0.95	0.95	0.92	0.95
	Missing	0.95	0.96	0.95	0.96	0.95
	Outlier	0.47	0.72	0.47	0.49	0.40
	Redundancy	0.96	0.95	0.96	0.96	0.95
Cancer	HD	1.00	1.00	1.00	0.57	0.81
	Imbalance	0.92	0.88	0.89	0.56	0.66
	Missing	0.71	0.74	0.68	0.58	0.56
	Outlier	0.42	0.53	0.40	0.08	0.24
	Redundancy	0.94	0.89	0.92	0.56	0.64
Diabetes	HD	0.83	0.83	0.75	0.59	0.67
	Imbalance	0.83	0.80	0.73	0.67	0.69
	Missing	0.75	0.77	0.69	0.59	0.66
	Outlier	0.36	0.59	0.33	0.31	0.27
	Redundancy	0.85	0.82	0.79	0.67	0.67
Fetal Health	HD	0.96	0.95	1.00	0.79	0.86
	Imbalance	0.98	0.95	1.00	0.58	0.78
	Missing	0.91	0.80	0.52	0.53	0.49
	Outlier	0.96	0.94	1.00	0.79	0.85
	Redundancy	0.96	0.93	1.00	0.72	0.85
Heart Failure	HD	1.00	0.98	1.00	0.64	0.76
	Imbalance	0.92	0.89	0.90	0.87	0.74
	Missing	0.84	0.84	0.72	0.67	0.68
	Outlier	0.45	0.71	0.40	0.31	0.28
	Redundancy	0.93	0.91	0.89	0.71	0.76
Lung Cancer	HD	0.92	1.00	0.88	0.46	0.61
	Imbalance	0.94	0.99	0.97	0.74	0.80
	Missing	0.85	0.90	0.65	0.29	0.72
	Outlier	0.73	0.76	0.58	0.36	0.68
	Redundancy	0.96	0.99	0.97	0.46	0.99
Mental Health	HD	0.49	0.71	0.45	0.44	0.52
	Imbalance	0.81	0.77	0.64	0.75	0.64
	Missing	0.51	0.75	0.48	0.44	0.55
	Outlier	0.50	0.76	0.46	0.37	0.49
	Redundancy	0.63	0.74	0.55	0.54	0.59

D. Model Evaluation

The model evaluation performed experimental studies for 8 datasets. The study compares the results for the actual best strategy based on model training with the best strategy recommended by the prototype model, as shown in Table III. The prototype model is able to recommend the actual active learning strategy for 7 out of 8 datasets. This indicates that the recommender model is accurate, correctly predicting the optimal active learning strategy for most datasets.

TABLE III. ACTUAL BEST STRATEGY VS BEST STRATEGY RECOMMENDED

Dataset	Strategy recommended by the prototype model	Actual best strategy based on experimental study
Alzheimer	Diversity sampling	Diversity Sampling
Asthma	QBC	QBC
Cancer	Uncertainty Sampling	Uncertainty sampling
Diabetes	QBC	QBC
Fetal Health	QBC	QBC
Heart Failure	Diversity sampling	QBC
Lung Cancer	QBC	QBC
Mental Health	QBC	QBC

Table IV shows the summary metrics for how many times each actual best strategy was correctly or incorrectly recommended for the 8 datasets.

TABLE IV. ACTIVE LEARNING RECOMMENDER METRICS

		Strategy recommended by the prototype model			
		UNC	QBC	DIVERSITY	TOTAL
Actual Best strategy	UNC	1			1
	QBC		5		5
	DIVERSITY		1	1	2
TOTAL		1	6	1	8

The study further evaluates the agreement between the prototype model recommended strategy and the actual best strategy, illustrated in Table IV matrix using Cohen's Kappa. The observed agreement, agreement of chance, and Cohen's Kappa formula were implemented as below:

Step 1: Calculate Observed Agreement (P_o)

Sum the diagonal elements (where both raters agree) and divide by the total number of observations [see Eq. (1)].

$$P_o = \frac{1+5+1}{8} = 0.875 \quad (1)$$

Step 2: Calculate Expected Agreement (Pe)

For each category, compute the product of the row and column totals, then sum these and divide by the total squared [see Eq. (2)]:

$$p_e = \frac{(1*1)+(5*6)+(2*1)}{8^2} = \frac{33}{64} \approx 0.5156 \quad (2)$$

Step 3: Compute Cohen's Kappa (κ)

Use the equation:

$$K = \frac{(P_O - P_e)}{(1-P_e)} = \frac{(0.875-0.5156)}{(1-0.5156)} = \frac{0.3594}{0.4844} \approx 0.742 \quad (3)$$

Based on the result obtained in Eq. (3), Cohen's Kappa value of 0.742 indicates a substantial agreement according to the commonly used interpretation scale, where values between 0.61 and 0.80 suggest substantial agreement, and values above 0.80 suggest almost perfect agreement. This implies that Cohen's Kappa result used in the study is reliable and not the result of random agreement. This level of agreement is important in the context of active learning strategies, which rely on iterative data selection and labeling. A high Kappa score ($\kappa = 0.742$) confirms the consistency and trustworthiness of the labeling process, ensuring that the performance of each strategy is based on reliable and ground truth data.

Given that medical datasets often suffer from data quality issues such as missing data and redundancy, reliable label consistency strengthens the validity of the training results. Specifically, the high accuracy observed for UNC and QBC methods (typically >0.90) across high-dimensional and imbalanced datasets is supported by the underlying agreement in labeling. Furthermore, it provides a sound justification for excluding cluster and density-based methods, whose poor and inconsistent performance may be exacerbated by sensitivity to labeling errors.

The obtained Cohen's Kappa results align with established findings on inter-rater reliability in medical and active learning settings. For example, McHugh et al. categorized kappa values between 0.61 and 0.80 as indicating substantial agreement, supporting the interpretation that a kappa of 0.742 reflects an acceptable level of reliability [29]. Similarly, Harmsen et al. demonstrated that Cohen's Kappa values exceeding 0.70 in active learning frameworks applied to medical literature screening correspond to reliable human annotation, even under noisy and incomplete label conditions [30]. These studies collectively affirm that the observed kappa value indicates robust inter-rater agreement and supports the reliability of the active learning evaluations conducted in this study.

V. CONCLUSION AND FUTURE WORK

This study examined the correlation between dataset characteristics and the efficacy of active learning strategies in predictive modelling. The research identified actual connections by analyzing critical data quality issues such as class imbalance, high dimensionality, missing values, redundancy, and outliers, which guide the optimal selection of active learning methods. A data-driven selection model was developed to identify suitable active learning strategies based on dataset characteristics, providing a systematic and evidence-based method to improve predictive accuracy and learning efficiency. Experimental validation across various healthcare datasets exhibited significant model reliability, with considerable concordance between anticipated and actual strategies. This study addresses a fundamental limitation in the AL literature by converting strategy selection from intuitive selections to a systematic, data-driven approach, hence enhancing the field's practical relevance in real-world context.

The study concentrated on foundational active learning strategies and specific healthcare datasets. Subsequent research should broaden the framework to include sophisticated, hybrid, and domain-specific methodologies, along with comprehensive performance metrics like interpretability, scalability, and computational efficiency. Incorporating multi-modal and real-world data, together with validation from domain experts, will significantly improve model generalizability. The suggested approach offers a pragmatic and theoretically substantiated contribution to the selection of adaptive, data-driven active learning strategies.

ACKNOWLEDGMENT

The authors would like to thank the Ministry of Higher Education Malaysia and Universiti Teknologi MARA for their financial support to this project under the Fundamental Research Grant Scheme (FRGS) Grant No. FRGS/1/2023/ICT01/UITM/02/4. We would also like to thank the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Selangor, Malaysia, for all the support.

REFERENCES

[1] P. Liu, L. Wang, R. Ranjan, G. He, and L. Zhao, "A Survey on Active Deep Learning: From Model Driven to Data Driven," ACM Comput Surv, vol. 54, no. 10, 2022, doi: 10.1145/3510414.

[2] D. Wu, R. Niu, M. Chinazzi, A. Vespignani, Y. A. Ma, and R. Yu, "Deep Bayesian Active Learning for Accelerating Stochastic Simulation," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2023. doi: 10.1145/3580305.3599300.

[3] U. Mahmood, Z. Fu, V. D. Calhoun, and S. Plis, "A deep learning model for data-driven discovery of functional connectivity," Algorithms, vol. 14, no. 3, 2021, doi: 10.3390/a14030075.

[4] M. Nashaat, A. Ghosh, J. Miller, and S. Quader, "Asterisk: Generating Large Training Datasets with Automatic Active Supervision," ACM/IMS Transactions on Data Science, vol. 1, no. 2, 2020.

[5] H. Huang, Q. Zhu, X. Zhu, and J. Zhang, "An Adaptive, Data-Driven Stacking Ensemble Learning Framework for the Short-Term Forecasting of Renewable Energy Generation," Energies (Basel), vol. 16, no. 4, 2023, doi: 10.3390/en16041963.

[6] L. Zhang and J. Wen, "Active learning strategy for high fidelity short-term data-driven building energy forecasting," Energy Build, vol. 244, 2021, doi: 10.1016/j.enbuild.2021.111026.

[7] X. Wang, X. Chi, Y. Song, and Z. Yang, "Active learning with label quality control," PeerJ Comput Sci, vol. 9, 2023, doi: 10.7717/peerj-cs.1480.

[8] A. Ashfaq, N. Cronin, and P. Müller, "Recent advances in machine learning for maximal oxygen uptake (VO2 max) prediction: A review," 2022. doi: 10.1016/j.imu.2022.100863.

[9] L. L. Sun and X. Z. Wang, "A survey on active learning strategy," in 2010 International Conference on Machine Learning and Cybernetics, ICMLC 2010, 2010. doi: 10.1109/ICMLC.2010.5581075.

[10] B. Settles, "Active Learning Literature Survey," Mach Learn, vol. 15, no. 2, 2010, doi: 10.1.1.167.4245.

[11] P. Ren et al., "A Survey of Deep Active Learning," 2022. doi: 10.1145/3472291.

[12] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri, and H. Salem, "Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction," Sensors, vol. 22, no. 3, 2022, doi: 10.3390/s22031184.

[13] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," 2021. doi: 10.1016/j.media.2021.102062.

[14] A. Alizadeh, P. Tavallali, M. R. Khosravi, and M. Singhal, "Survey on Recent Active Learning Methods for Deep Learning," 2021. doi: 10.1007/978-3-030-69984-0_43.

[15] L. Riyaz, M. A. Butt, M. Zaman, and O. Ayob, "Heart Disease Prediction Using Machine Learning Techniques: A Quantitative Review," 2022. doi: 10.1007/978-981-16-3071-2_8.

[16] J. Chen, H. Huang, A. G. Cohn, D. Zhang, and M. Zhou, "Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning," Int J Min Sci Technol, vol. 32, no. 2, 2022, doi: 10.1016/j.ijmst.2021.08.004.

[17] D. Lamba, W. H. Hsu, and M. Alsadhan, "Predictive analytics and machine learning for medical informatics: A survey of tasks and techniques," in Machine Learning, Big Data, and IoT for Medical Informatics, 2021. doi: 10.1016/B978-0-12-821777-1.00023-9.

[18] Z. Pan, P. Soong, and S. Rafatirad, "Ontology-Driven Scientific Literature Classification Using Clustering and Self-supervised Learning," in Lecture Notes on Data Engineering and Communications Technologies, vol. 137, 2023. doi: 10.1007/978-981-19-2600-6_10.

[19] B. R. Devi, U. Sivaji, T. Swetha, J. Avanija, A. Suresh, and K. R. Madhavi, "Advanced Cardiovascular Disease Prediction: A Comparative Analysis of Ensemble Stacking and Deep Neural Networks," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 6, 2024.

[20] K. Kanwar, S. Sonia, K. Aggarwal, D. D. Solomon, and K. Polat, "Triple Voting: Hybrid Cardiovascular Diseases Prediction Model," International Journal of Applied Decision Sciences, vol. 1, no. 1, 2025, doi: 10.1504/ijads.2025.10061195.

[21] J. Chang, J. Kim, B. T. Zhang, M. A. Pitt, and J. I. Myung, "Data-driven experimental design and model development using Gaussian process with active learning," Cogn Psychol, vol. 125, 2021, doi: 10.1016/j.cogpsych.2020.101360.

[22] M. Bernhardt et al., "Active label cleaning for improved dataset quality under resource constraints," Nat Commun, vol. 13, no. 1, 2022, doi: 10.1038/s41467-022-28818-3.

[23] D. Mahapatra, A. Poellinger, L. Shao, and M. Reyes, "Interpretability-Driven Sample Selection Using Self Supervised Learning for Disease Classification and Segmentation," IEEE Trans Med Imaging, vol. 40, no. 10, 2021, doi: 10.1109/TMI.2021.3061724.

[24] E. Tasci, Y. Zhuge, K. Camphausen, and A. V. Krauze, "Bias and Class Imbalance in Oncologic Data—Towards Inclusive and Transferrable AI in Large Scale Oncology Data Sets," 2022. doi: 10.3390/cancers14122897.

[25] L. Wang et al., "GAN-Based Dual Active Learning for Nosocomial Infection Detection," IEEE Trans Netw Sci Eng, vol. 9, no. 5, 2022, doi: 10.1109/TNSE.2021.3100322.

[26] C. J. Arthurs and A. P. King, "Active training of physics-informed neural networks to aggregate and interpolate parametric solutions to the Navier-Stokes equations," J Comput Phys, vol. 438, 2021, doi: 10.1016/j.jcp.2021.110364.

[27] L. El bouny, M. Khalil, and A. Adib, "An End-to-End Multi-Level Wavelet Convolutional Neural Networks for heart diseases diagnosis," Neurocomputing, vol. 417, 2020, doi: 10.1016/j.neucom.2020.07.056.

[28] Y. Han, B. Tang, and L. Deng, "Multi-level wavelet packet fusion in dynamic ensemble convolutional neural network for fault diagnosis," Measurement (Lond), vol. 127, 2018, doi: 10.1016/j.measurement.2018.05.098.

[29] M. L. McHugh, "Interrater reliability: The kappa statistic," Biochem Med (Zagreb), vol. 22, no. 3, 2012, doi: 10.11613/bm.2012.031.

[30] S. M. Salhout, "Machine learning in healthcare strategic management: a systematic literature review," 2024. doi: 10.1108/AGJSR-06-2023-0252.