Journal of Computer Applications and Information Technology
Vol. 1, Issue. 3, Sep 2025, pp. 60~72
ISSN: 3107-6424

Enhancing Predictive Accuracy Using Ensemble Machine Learning Models: A Data-Centric Approach

Ushik Shrestha[1], Swetha Indudhar Goudar[2]
[1]School of Science and Technology, RMIT University Vietnam, Hanoi, Vietnam.
[2]Professor & Dean Research and Development, KLS Gogte Institute of Technology, Belagavi, Karnataka.
E-mail: ushik.shrestha@rmit.edu.vn[1], sigoudar@git.edu[2]

Article Info

Article History:

Received Jul 05, 2025
Revised Aug 03, 2025
Accepted Sep 02, 2025

Keywords:

Data-Centric Approach
Machine Learning
Artificial Intelligence
Big data analytics
Prediction Accuracy

ABSTRACT

Due to the abundance of data and processing capacity, artificial intelligence (AI) has developed and is now linked to deep learning. In order to optimize the efficiency without changing the underlying data, academics have traditionally taken a Data-centric approach, concentrating on creating innovative models and algorithms. The Data Centric Approach, also called the Data Oriented model, was created as a result of notable AI expert Andrew Ng's current emphasis on improved (quality) data rather than better methods. In the field of ML, the shift from a model-oriented to a data-oriented model has accelerated. Notwithstanding its potential, the Data-Centric Approach has a number of obstacles to overcome, such as (a) producing high-quality data, (b) protecting data privacy, and (c) resolving biases to make datasets equitable. There hasn't been much work done lately to prepare high-quality data. By concentrating on producing high-quality data using techniques like data augmentation, multistage hashing to remove duplicate instances, detecting and correcting noisy labels, and confident learning, our study seeks to close this gap. Our study presents an Enhancing Predictive Accuracy Using Ensemble Machine Learning Models: A Data-Centric Approach (EPAE-MLDCA). The EPAE-MLDCA continuously beat the data model centric model according to a comparative performance study. This research shows how the EPAE-MLDCA approach may be further investigated and used in a variety of fields, including entertainment, finance, healthcare, and education, where high-quality data could greatly improve performance. We discovered that compared to data-centric techniques, the EPAE-MLDCA approach provided a greater accuracy.

Corresponding Author:

Ushik Shrestha,
School of Science and Technology,
RMIT University Vietnam, Hanoi, Vietnam.
E-mail: ushik.shrestha@rmit.edu.vn

1. INTRODUCTION

Data is and has always been at the heart of machine learning (ML). However, the advent of powerful push-button models has only recently caused data science teams to shift their attention to data. This approach, called data-centric ML, involves building intelligent models with high-quality data, with a focus on making sure the data effectively conveys the knowledge the AI needs to acquire [1]. It also involves working together and iterating on the information required to programmatically create AI systems. Iterations of custom model architecture, feature engineering, and algorithm design have historically been the focus of data science and ML teams when creating models. The majority of teams focused on the model, and they viewed the data as static elements [2]. But as models have grown increasingly complex and push-button, AI teams have come to understand that data iteration is equally, if not more, crucial to the successful and effective development and implementation of high-accuracy models [3].

In recent years, ML models have grown more complex and opaque, requiring much more training data. Data has also developed into a practical interface for collaborating with subject matter specialists and turning their knowledge into software. Lastly, a greater degree of model accuracy than was previously possible with simply data-centric approaches is made possible by data-centric ML. Three types of data sources can be categorized: semi-structured, unstructured, and structured data. About 80 percent of all data worldwide is unstructured, whereas only 20 percent is structured. Complexity can arise from large data volumes [4]. As a result, fewer people are using basic data analysis techniques. The amount of unstructured and semi-structured data is just too large to handle. Big data analytics (BDA) is a phrase that researchers have created to refer to [5] complex and big datasets. “The term 'big data' describes the growing volume of both structured and unstructured data. Simple data management tools cannot handle relational databases. In order to manage big and complicated datasets, BDA compiles a variety of tools and approaches [6].

A wide definition of "big data" has been developed by researchers using a number of criteria. First, three attributes of big data were identified by researchers: variety, speed, and quantity. In order to properly describe high-quality data, researchers started adding more data characteristics as big data became more complicated. These characteristics are referred to by the V of big data. Lastly, depending on the challenges of handling massive volumes of data, some scholars have defined up to features of big data. In the following sections, we discussed the ten V's of big data [7]. Continuous enhancement of data quality is the aim of data-centric ML. It undergoes numerous processes in the ML lifecycle after being collected and generated. Following data cleaning, the data undergoes labeling and augmentation before analysis. An essential component of data-centric ML is preprocessing. Numerous experiments were carried out using the data generated during the model training phase in order to improve the accuracy [8]. Two major elements have contributed to the extraordinary development of the field of AI: the growing amount of data and the ever-increasing computer power. Multiple-layered neural networks used in DL, a branch of AI, have demonstrated an amazing ability to solve challenging issues.

A vast amount of data, produced by a variety of sources such as sensors, social media, e-commerce, etc., is necessary for deep learning to be successful. Furthermore, with the development of powerful GPUs, Deep Learning models have an insatiable appetite for computational resources. Researchers have been using the Model Centric Approach to tackle complicated problems over the past

few years. Without altering the current data, the Model Centric Approach seeks to develop advanced models and algorithms and improve model performance through hyperparameter tuning. However, a recent change suggests that both the sheer volume of data and high-quality data help deep models perform better. While using AI as an auxiliary tool rather than a replacement for human expertise, this approach acknowledges the intrinsic significance of human engagement in complicated processes [9]. The data-centric strategy poses a number of difficulties, including (a) producing high-quality data, (b) protecting data security and privacy, and (c) resolving biases in datasets for fairness. Notably, prominent AI specialist Andrew Ng has recently emphasized the importance of the data centric model, saying that the focus should shift from constantly improving AI models to prioritizing the improvement of the underlying data. The data-centric approach has several challenges, such as (a) generating high-quality data, (b) safeguarding data security and privacy, and (c) eliminating biases in datasets for equity.

In computer science, artificial intelligence (AI) is a rapidly expanding field. In research, artificial intelligence aims to increase accuracy while using the fewest resources possible. Accuracy for a ML model will rise with time; the more training epochs (the longer the training period), the better the accuracy. Its accuracy will eventually stop increasing, though, which presents a challenge for researchers: how to keep improving the model when model training is no longer successful. The data and the model (or neural network structure) are the two components that make up a ML model. In one example, the model is referred to as the body and the data as the nourishment. Since "we are what we eat," the model (or body) will be "healthy" if the data is good, but "unhealthy" if the data is bad. Therefore, concentrating on the quality of the data is one way to address the issue of model accuracy. High-quality data will help a model perform better. As a result, data preparation has been the focus of ML

Stated differently, a recent development in ML is data-centric ML. High quality data labeling, or the process of giving the data one or more labels, is necessary for data-centric ML in addition to sophisticated preprocessing methods. Both ML operations (MLOps) and data providers must put out effort in this regard. A subset of AI research known as "data-centric AI" focuses on optimizing models by adjusting cost functions and hyperparameters. Another subgroup of AI that concentrates on preparing data to enhance the quality of the data that will ultimately be fed into models is called data-centric ML. While data-centric AI may tolerate incorrect data labels, it demands consistency in the data. The data-centric strategy invests in data quality technologies to clean noisy data, whereas the data-centric approach optimizes the model to handle noisy data. A data-centric strategy iterates the quality of the data, whereas a data-centric approach iteratively improves the model. When compared to data-centric AI, data-centric AI has numerous benefits. First, while data advancements continue to demonstrate efficiency, model advancements are presumed to achieve a benchmark. Second, because data is simpler for domain specialists to understand than mathematical formulae, data-centric approaches enable more of them to contribute.

We postulated that data-centric AI will enhance a ML model's performance in light of these benefits. Three ML models—two Data-centric—were employed in total to test this hypothesis. The data-centric approach concentrated on improving the accuracy of the data by refining it using sophisticated techniques, whereas the two data-centric approaches concentrated on the neural network structures and the number of training epochs (iterations). The EPAE-MLDCA continuously beat the data centric approach according to comparative performance. This research shows how the EPAE-

MLDCA approach may be further investigated and used in a variety of fields, including entertainment, finance, education, and healthcare, where high-quality data could greatly improve performance. We discovered that compared to data-centric techniques, the EPAE-MLDCA approach provided a greater accuracy.

2. RELATED WORKS

Rahal et al. [11] provide a clever framework for data-centric evaluation that may find high-quality data and enhance an ML system's performance. To differentiate between high- and low-quality data, the suggested system blends unsupervised learning with the curation of quality assessments. In order to be implemented across a range of domains and applications, the framework is made to incorporate adaptable and general-purpose techniques. We applied the developed framework to a real-time application from the domain of analytical chemistry, testing it on three datasets of anti-sense oligonucleotides in order to verify the results. To determine the pertinent quality metrics and assess the framework's results, a domain expert is consulted. The findings demonstrate how the quality-centric data evaluation framework pinpoints the qualities of high-quality data that direct the execution of productive lab tests and, as a result, enhance the ML system's performance.

Sánchez-Marqués et al. [12] examine a data-driven ML strategy to ascertain its possible advantages in glioma grade prediction. To give a comprehensive view of model performance, we present six performance metrics. Experiments on a low-imbalanced data set comprising clinical variables and molecular biomarkers show that standardization and oversizing the minority class improve the prediction performance of two classifier ensembles and four well-known ML models. Additionally, the trials demonstrate that three of the four conventional prediction models are significantly outperformed by the two classifier ensembles. Additionally, utilizing four feature ranking algorithms, we perform a thorough descriptive analysis of the glioma data set in order to find the most useful qualities and pertinent statistical characteristics.

Khan et al. [13] examined the possibility of various data-centric approaches to enhance the learning of ML models, particularly NNs, on fewer frequent failures with such unbalanced training datasets. With an enhancement of up to 5.5% in F1-score seen on less frequent failures compared to a baseline NN (i.e., without any data-centric or data-centric treatment), the results obtained for failure identification indicate that data-centric models tend to operate more effectively in terms of classification accuracy. A suitable strategy should be selected depending on the the available computational resources and intended classification accuracy, as some data-centric approaches may also come with a large amount of additional computational complexity.

Mehta and Patnaik [14] intend to lessen the impact of the two main problems—data imbalance and the large dimensionality of the defect datasets—that arise during defect prediction. Feature selection approaches are used in this study to analyze a variety of software metrics. Because of the unbalanced nature of the datasets employed, the suggested model combines partial least squares (PLS) regression and RFE for dimension reduction, which is then integrated with the SMOTE. Comparing the methods utilized in the study, it was found that the XGBoost and Stacking Ensemble techniques produced the best outcomes for all datasets, with defect prediction accuracy greater.

3. THE PROPOSED MODEL

ML technique enables computers to learn from data without explicit programming. When given data that it did not encounter during the training phase [15], an efficient ML system can produce accurate predictions for the task for which it was created. Iteratively fitting different models to processed raw data is typically necessary for the construction of such a system in order to find the optimal ML model. The ML community currently uses a data-centric ML model design, which is focused on finding the optimal model using a predetermined set of attributes. A structure of a data-centric ML approach is shown in Figure 1. This approach's primary feature is that every step before model training is a one-time operation. They are only run once at the start of the ML model design process. This methodology's core component is hyperparameter-tuned model training. The primary flaw in the data-centric approach is that it fails to consider the fact that the accuracy of the final model is dependent on the caliber of the data that is fed into it, as well as the model type and hyperparameter settings.

Data frequently has traits that contribute to overfitting, which is even more concerning, but do not increase precision. Additionally, it could be difficult to find inaccurate data in specific aspects for larger data sets. The resulting model's accuracy is reduced in both cases, and even with considerable hyperparameter adjustment, it cannot be made more accurate. The most popular method for addressing this problem is to increase data collection in order to make up for any potential shortcomings. However, it is not practical to gather vast amounts of data in many fields. As a result, it's critical to develop and investigate ways to enhance the data we now have. Data-centric ML refers to methods that try to enhance the data that goes into ML models in order to improve their performance. It goes without saying that focusing only on data and not looking for the optimal model would probably not increase the end system's precision.

Nevertheless, enhancing the conventional data-centric method by carefully selecting and enhancing the quality of the features to be taken into account tends to both decrease the computing time of training procedures and improve the final ML model's precision.

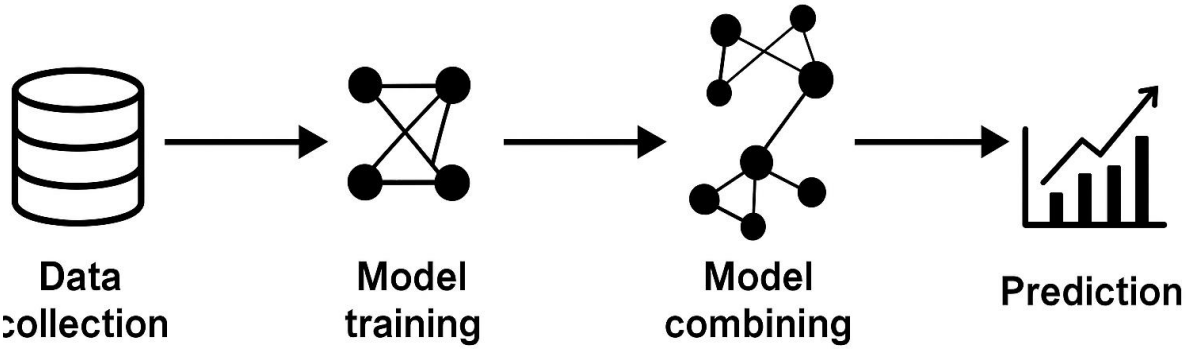

Figure 1. Structure of Data-Centric ML approach

3.1. Data-centric ML approaches

Basic data pretreatment techniques were necessary to accurately feed the data into the models, even though the AI approaches were not data-centric. Initially, all superfluous columns were removed, such as ID and Unnamed. The diagnosis results were then transformed into numbers using

pandas.get_dummies function because ML algorithms can only comprehend numerical inputs [16]. More precisely, "benign" and "malignant" were changed to 1 and 0, respectively. Following data preprocessing, 569 samples (rows) had 30 characteristics (columns). The train_test_split function in Scikit-learn was then used to randomly divide the data using an 80/20 training/testing paradigm. Artificial Neural Network (ANN)-based deep learning works were developed. Three layers of neurons make up an ANN: the input layer, the hidden layer or layers, and the output layer. Neuronal connections in two successive layers were assigned a weight based on the relative relevance of the input information. The data was then subjected to an activation function in order to normalize the neuron's output.

A cost function that showed the discrepancy between the expected and actual outputs was generated by iterating through the dataset. After each iteration (also known as an epoch), gradient descent was used to adjust the weights between neurons in order to minimize this cost function. The fact that every sample was classified as either benign or malignant suggested that there was a binary classification issue.

Following data preparation, there were 30 features; hence, the input layer had 30 nodes. There were 50, 30, and 20 nodes in each of the three concealed layers. There was only one node in the output layer. The optimizer Adam and the loss function Binary Cross-Entropy were used to compile the model. The final accuracy was then obtained after 60 epochs (iterations) of training. With seven layers, the NN of the data-centric approach 2 was more intricate than that of the data-centric approach 1. There were thirty nodes in the input layer.

3.2. Data-centric ML approach

The data-centric ML strategy prioritized data pretreatment above model construction and training. In addition to the processes of the first technique, data pretreatment would include outlier detection, feature engineering, and dataset balance. The technique of altering a dataset's characteristics to enhance ML model training is known as feature engineering. The four processes of feature engineering are feature extraction, feature transformation (also known as feature scaling), feature creation, and feature selection. Relationships between characteristics are typically present in datasets (e.g., height may be related to weight). We may employ feature creation, which is the process of developing new features by utilizing interactions between preexisting ones, to capitalize on these relationships [16].

The equation X3 = X1*X2 represents a straightforward two-way interaction, where X3 is the interaction between two distinct features in the dataset, X1 and X2. The dataset contained 30 features prior to feature generation. There would be 435 (30*29/2) interactions between these 30 characteristics. Following this stage, there were 465 features in total (435+30). The data's feature range is normalized by feature scaling. The gradient descent converges more smoothly with feature scaling, which enhances the model and shortens the training period. The standardization technique, which is used in feature scaling, centers the values around the mean with a unit standard deviation. The 465 features remain unchanged as a result of standardization. Training time is greatly decreased by feature extraction, which eliminates redundant data from the data set. This stage makes use of principal component analysis, or PCA. Using the PCA technique, a dataset with several attributes is broken down into primary components that represent the variation underlying the data. Finding the linear

feature combination that maximizes variance while maintaining zero correlation with the previously determined principal components is how each principal component is determined.

Following PCA, there were 10 features instead of 465. A subset of pertinent features is chosen for use in model creation through feature selection. Dimensionality reduction is not the same as feature selection. Both techniques reduce the amount of attributes in the dataset, but feature selection adds and removes features without changing them, while dimensionality reduction generates new attribute groups. Eight of those ten traits were chosen and kept in the dataset. There were 173 malignant and 282 benign samples in this collection. It was crucial to balance the dataset because this imbalance might have led to model bias. SMOTE replicates minority class cases, increasing them at random. Following the use of SMOTE, there were 282 samples with as many benign as malignant characteristics. The data-centric approach and data-centric methods used the same methodology for dataset separation. The data-centric approach's five-layer neural network was similar to that of the data-centric approach 1's. There were thirty nodes in the input layer. There were 50, 30, and 20 nodes in the concealed layers. There was only one node in the output layer.

3.2.1 Data Pre-processing

This stage involves applying and assessing several preprocessing techniques on the underlying model. Important decisions that are typically made quite naturally are frequently incorporated into data preprocessing. These options include designing an effective target variable, creating additional features, developing a strategy for handling missing values, and implementing data cleaning techniques. Every solution under consideration is assessed using a basic model in the suggested data-centric methodology. Thus, the optimal set of tactics is selected as a row data preparation procedure. The pipeline's subsequent stage subsequently makes use of the preprocessed data.

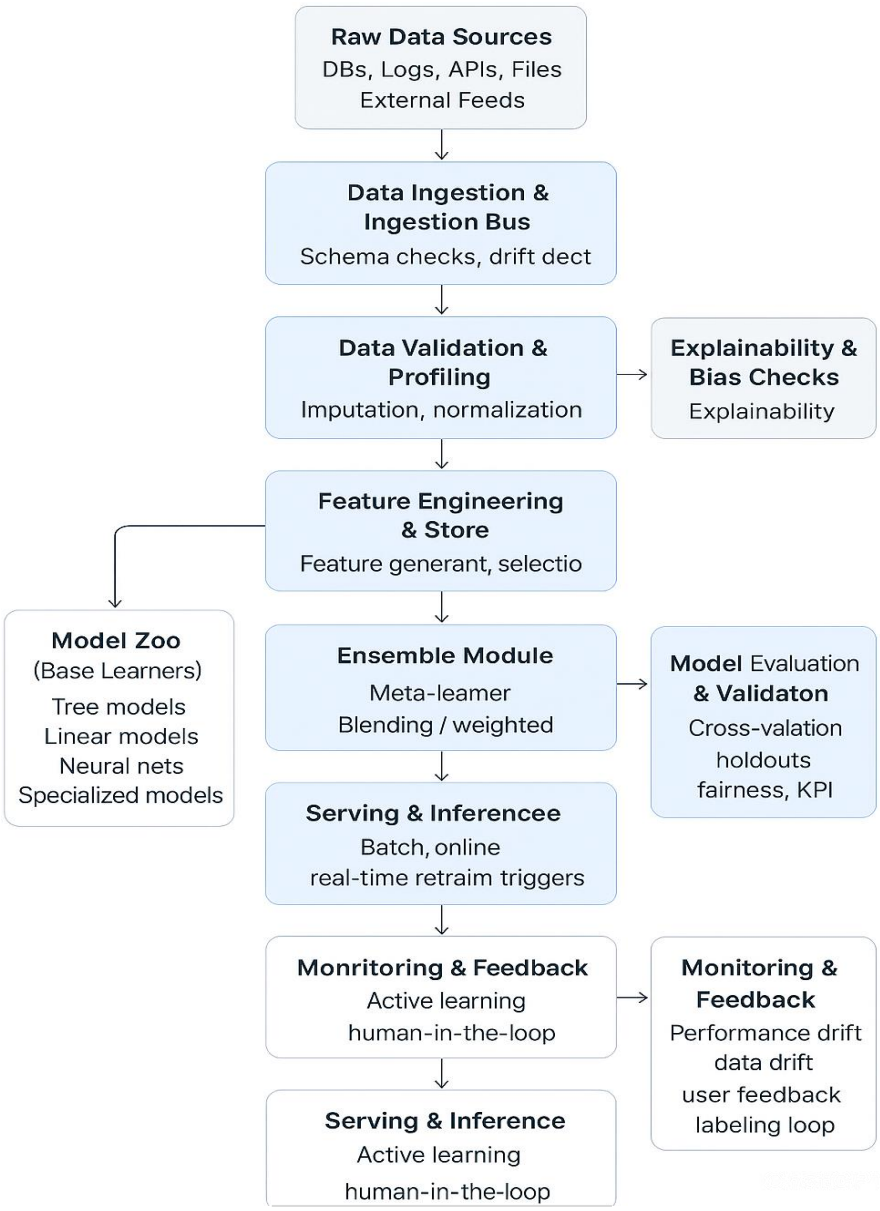

Figure 2. Overall architecture of the EPAE-MLDCA approach

3.2.2. Feature Selection

Real-world data frequently has a broad range of qualities, not all of which are instructive for the process they are intended to depict. In addition to increasing computation time, the presence of non-informative features causes the final models to overfit. As a result, choosing the best features from the data set can significantly impact how well ML models perform. Manual selection based on domain expertise is the most used feature selection technique. Nevertheless, because this value has not yet been determined, certain features with high predictive value may be left out of the analysis when employing this method.

Furthermore, acquiring in-depth domain knowledge can be challenging for certain ML applications.

We can apply several feature selection strategies to the preprocessed data set and assess their effectiveness using the suggested data-centric methodology. The model training stage then selects the parameter that yields the highest precision on the validation set.

3.2.3. Model Training

Similar to the data-centric method to ML, the model training stage is carried out. The data set is fitted to a selected ML model using different hyperparameter values. The final model is selected based on which configuration performs the best on the validation set. The test set, which was not utilized in the pipeline, is used to report the final model's precision.

4. EXPERIMENTAL RESULTS

We compared data-centric methods with EPAE-MLDCA approach. While the data-centric strategy had sophisticated data preparation procedures and model training for 60 epochs, the two data-centric approaches included. There were ten test runs, or repetitions. Following 10 test runs and model trainings, the data-centric strategy had an accuracy of 96.6%, while the data-centric approaches 1 and 2 had an accuracy of 90.8% and 89.6%, respectively (Table 1).

With p-values of 0.00002 and 0.0008, both less than 0.05, our outcomes confirm that the EPAE-MLDCA strategy outperforms data-centric approaches 1 and 2, respectively. Confusion matrices were used to characterize how well a classification model performed on a set of test data. Out of 143 predictions, 4 false positives, 56 true negatives, 84 genuine positives, and 4 false negatives were predicted by the data-centric approach 1 (Table 2). Out of 116 predictions, 8 false positives, 56 true negatives, 79 genuine positives and 2 false negatives, were predicted by the data-centric approach 2 (Table 3).

Out of 150 predictions, 82 genuine positives 4 false positives, 52 true negatives, and 4 false negatives, were predicted using the data-centric approach (Table 4). Both data-centric approaches predicted an average of four false negatives out of 143 data points (3% rate) in this specific breast cancer prediction problem. This indicates that the models failed to identify the patient's cancer 3% of the time, which could lead to the patient not receiving treatment in a clinical context (Table 2-3).

A considerably superior outcome was obtained in a clinical scenario, as evidenced by the data-centric model prediction of 2 false negatives out of 160 data points (1.4% rate) in comparison to this figure (Table 4). Three models' accuracy gains over the course of training were also taken into account (Figure 1). Tensorboard, a ML visualization and toolkit, was used to do this.

The models only obtained accuracies of less than 60% for data-centric approaches 1 and 2 in the first epoch. The models achieved 96% and 95% accuracy after 60 epochs of training. Initially, the model's accuracy for the data-centric method was over 80%. The model achieved 98% accuracy after only 20 epochs of training. It is evident that improving data greatly improves model performance right from the start of training.

Table 1. Accuracy of EPAE-MLDCA approach with other data-centric approaches

	Data Centric Approach 1	**Data Centric Approach 2**	**EPAE-MLDCA Approach**
Test run #1-10	0.78	0.86	0.96
	0.87	0.90	0.89
	0.8	0.85	0.94
	0.7	0.92	0.93
	0.83	0.79	0.89
	0.7	0.80	0.97
	0.80	0.9	0.97
	0.76	0.91	0.89
	0.74	0.95	0.96
	0.89	0.84	0.97
Mean	0.802	0.783	0.986
Standard Deviation	0.000023412	0.00016556	5.4698976
Variance	0.000010102	0.003543333	8.2211121

Table 2. Confusion matrix of data-centric model 1

	Predicted: No	**Predicted: Yes**
Actual: No	**56**	**8**
Actual: Yes	**2**	**79**

Table 3. Confusion matrix of data-centric model 2

	Predicted: No	**Predicted: Yes**
Actual: No	**52**	**13**
Actual: Yes	**2**	**82**

Table 4. Confusion matrix of EPAE-MLDCA approach

	Predicted: No	**Predicted: Yes**
Actual: No	**68**	**4**
Actual: Yes	**2**	**69**

5. DISCUSSIONS

According to our findings, EPAE-MLDCA has the potential to significantly enhance ML in terms of accuracy and processing outcomes (number of epochs). Finally, the data-centric technique only required 20 epochs to achieve 98% accuracy, whereas the two data-centric approaches produced approximately 95% accuracy after 60 epochs. In order to determine whether the findings of the study were generalizable and not model-specific, we decided to employ two data-centric methodologies. Figure 3 show the accuracy of EPAE-MLDCA with other models. Additionally, as results can vary from one study to the next, ten test runs were employed to increase the research's statistical robustness.

When three approaches are compared on a dataset, the data-centric strategy with many sophisticated data pretreatment techniques performs the best, demonstrating the significance of data refinement.

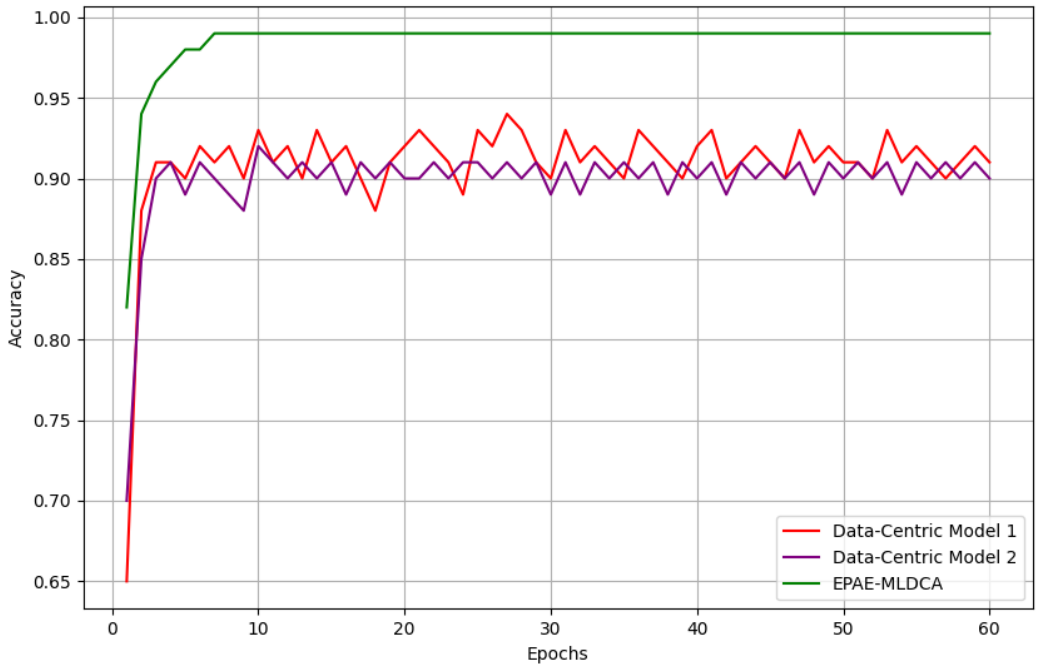

Figure 3. Accuracy of EPAE-MLDCA with other models

The shift to data-centric ML can have a number of effects. First, 80% of ML labor involves data preparation, and 20% involves model development. This is because, being the most efficient step in the process, time should be spent on data preparation rather than model training, which uses more time and computing resources but produces lower accuracy. Second, ML models are becoming commoditized as selecting and creating a model no longer constitutes the main component. This implies that pip install, a single line of code, may easily accomplish deep learning. To make the process of deploying ML easier, there are a number of ML API services (such as Bonsai, AWS, Azure, GCE, and Clarifai) and AutoML tools (such as Auto-sklearn, H20.ai, and Auto-Keras). This promotes the use of ML across a range of industries, including defense, manufacturing, retail, transportation, and healthcare. Thirdly, a variety of advances regarding data management, collection, and labeling can be produced because data becomes the focal point of ML, leading to high-growth startups.

6. CONCLUSION

Our study presents an Enhancing Predictive Accuracy Using Ensemble Machine Learning Models: A Data-Centric Approach (EPAE-MLDCA). The Data Centric Approach continuously beat the Model Centric Approach according to a comparative performance study. This research shows how the EPAE-MLDCA approach may be further investigated and used in a variety of fields, including entertainment, healthcare, finance, and education, where high-quality data could greatly improve performance. We discovered that compared to data-centric techniques, the EPAE-MLDCA approach provided a greater accuracy. Beyond the use case examined in this paper, the suggested EPAE-

MLDCA methodology can be applied to optimize the performance of ML approaches. It can be applied to a wide range of data-oriented tactics in addition to being generalizable across different domains.

REFERENCES

[1] Sánchez-Marqués, R., García, V. and Sánchez, J.S., 2024. A data-centric machine learning approach to improve prediction of glioma grades using low-imbalance TCGA data. Scientific Reports, 14(1), p.17195.

[2] Pan, I., Mason, L.R. and Matar, O.K., 2022. Data-centric Engineering: integrating simulation, machine learning and statistics. Challenges and opportunities. Chemical Engineering Science, 249, p.117271.

[3] Zhang, T., Wang, D. and Lu, Y., 2024. A data-centric strategy to improve performance of automatic pavement defects detection. Automation in Construction, 160, p.105334.

[4] Rahal, M., Ahmed, B.S., Szabados, G., Fornstedt, T. and Samuelsson, J., 2025. Enhancing machine learning performance through intelligent data quality assessment: An unsupervised data-centric framework. Heliyon, 11(4).

[5] Tschalzev, A., Marton, S., Lüdtke, S., Bartelt, C. and Stuckenschmidt, H., 2024. A data-centric perspective on evaluating machine learning models for tabular data. Advances in Neural Information Processing Systems, 37, pp.95896-95930.

[6] Bhatt, N., Bhatt, N., Prajapati, P., Sorathiya, V., Alshathri, S. and El-Shafai, W., 2024. A data-centric approach to improve performance of deep learning models. Scientific Reports, 14(1), p.22329.

[7] Singh, P., 2023. Systematic review of data-centric approaches in artificial intelligence and machine learning. Data Science and Management, 6(3), pp.144-157.

[8] Bhowmik, P. and Partha, A.S., 2021. A data-centric approach to improve machine learning model's performance in production. Int. J. Eng. Adv. Technol.(IJEAT), 11, pp.240-243.

[9] Westermann, H., Šavelka, J., Walker, V.R., Ashley, K.D. and Benyekhlef, K., 2021. Data-centric machine learning: Improving model performance and understanding through dataset analysis. In Legal Knowledge and Information Systems (pp. 54-57). IOS Press.

[10] La, H. and La, K., 2023. Comparing data-centric and data-centric approaches to determine the efficiency of data-centric AI. Journal Of Emerging Investigators.

[11] Rahal, M., Ahmed, B.S., Szabados, G., Fornstedt, T. and Samuelsson, J., 2025. Enhancing machine learning performance through intelligent data quality assessment: An unsupervised data-centric framework. Heliyon, 11(4).

[12] Sánchez-Marqués, R., García, V. and Sánchez, J.S., 2024. A data-centric machine learning approach to improve prediction of glioma grades using low-imbalance TCGA data. Scientific Reports, 14(1), p.17195.

[13] Khan, L.Z., Pedro, J., Costa, N., Sgambelluri, A., Napoli, A. and Sambo, N., 2024. Model and data-centric machine learning algorithms to address data scarcity for failure identification. Journal of Optical Communications and Networking, 16(3), pp.369-381.

[14] Mehta, S. and Patnaik, K.S., 2021. Improved prediction of software defects using ensemble machine learning techniques. Neural Computing and Applications, 33(16), pp.10551-10562

[15] Babbar, H., Rani, S., Singh, A., Abd-Elnaby, M. and Choi, B.J., 2021. Cloud based smart city services for industrial internet of things in software-defined networking. Sustainability, 13(16), p.8910.

[16] Rahouti, M., Xiong, K. and Xin, Y., 2020. Secure software-defined networking communication systems for smart cities: Current status, challenges, and trends. Ieee Access, 9, pp.12083-12113.

[17] Alshahrani, M.M., 2023. A Secure and intelligent software-defined networking framework for future smart cities to prevent DDoS Attack. Applied Sciences, 13(17), p.9822.