OPEN ACCESS

Global Journal of Research in Engineering & Computer Sciences
ISSN: 2583-2727 (Online)
Volume 05| Issue 02 | March-April | 2025
Journal homepage: https://gjrpublication.com/gjrecs/

Research Article

Selecting and designing appropriate databases for deep learning models

*Reza Yavari[1], Mostafa Boroumandzadeh[2]

[1] Postgraduated Student, Department of Computer Engineering and Information Technology, Payam Noor University, Tehran, Iran.
[2] Assistant Professor, Department of Computer Engineering and Information Technology, Payam Noor University, Tehran, Iran. ORCID: 0000-0002-8451-5364

DOI: 10.5281/zenodo.15298240 **Submission Date: 24 March 2025 | Published Date: 28 April 2025**

Corresponding author:** ***Reza Yavari

Postgraduated Student, Department of Computer Engineering and Information Technology, Payam Noor University, Tehran, Iran.

Abstract

In deep learning projects, selecting and designing the appropriate database can have a significant impact on the performance and efficiency of models. Given the complexity and diversity of data needs in this field, identifying the factors that affect this selection is of particular importance. The aim of this study is to investigate and analyze the factors affecting the selection and design of appropriate databases for deep learning models. This research used a purposive sampling method to select 30 experts and researchers active in the fields of artificial intelligence and deep learning. The data collection tool included structured questionnaires and semi-structured interviews, and the data were analyzed using descriptive and inferential statistical methods and thematic analysis. The findings show that the level of education and practical experience of individuals, type of organization, participation in international projects, scalability needs, and the importance of data security affect the selection of the type of database. Individuals with higher education and more experience tend to use NoSQL databases. The results show that choosing the right database requires attention to a set of individual and organizational factors that can lead to improved performance and efficiency of deep learning models.

Keywords: *Deep learning, database, NoSQL, scalability, data security, database selection, artificial intelligence.*

1. Introduction

Selecting and designing appropriate databases for deep learning models is a critical aspect of the success of AI projects. Databases serve as the backbone of machine learning processes, and their quality and structure can have a direct impact on the performance of deep learning models (Sami, Richard, Gauchard, Estève, & Rossi, 2022). In recent years, the exponential growth of data and technological advances in data processing have enabled the use of deep learning models in many areas. However, the wrong choice of database can lead to inaccurate and unreliable results. For this reason, the process of selecting and designing a database must be done carefully, taking into account the specific needs of each project (Yang & Gao, 2022).

In a study conducted by Wang et al. (2022), researchers concluded that data quality in terms of accuracy, completeness, and diversity can have a significant impact on the performance of deep learning models. They showed that using incomplete or incorrect data can lead to a decrease in the accuracy of the model, and as a result, affect the final results of the project (Zou, You, Wang, Wen, & Jia, 2022). In addition to the quality of the data, its volume and diversity are also of great importance. Deep learning models require a large amount of diverse data to learn complex patterns. This data must be designed in a way that covers all aspects of the problem at hand and allows the model to identify different patterns (Li & Zhou, 2022).

In another study conducted by Saputra and colleagues in 2024, researchers investigated the impact of database design on the performance of deep learning models. They showed that proper database design can help improve model performance and reduce training time. This study emphasizes that database design should be done according to the structure and needs of the model (Saputra, Riza, Setiawan, & Hamidah, 2024). Another important aspect in choosing a database is data accessibility and legal issues related to them. Many databases have access restrictions for various reasons, such as privacy or intellectual property rights. This can create challenges for researchers in using appropriate data (Saleh, Jhanjhi, Abdullah, & Saher, 2022).

Also, the tools and technologies used to store and manage data play an important role in the database selection and design process. Choosing the right tools can help improve the efficiency and speed of learning processes, and as a result, yield better results (Wang et al., 2023). Another important aspect is the issue of data preprocessing. Raw data usually requires preprocessing to make it suitable for use in deep learning models. This process involves cleaning the data, normalizing it, and converting it into a format that the model can use (Lim et al., 2020).

Selecting and designing appropriate databases should be done with the specific goals and needs of the project in mind. Each project may have its own data needs, and therefore, a generic approach will not be suitable for all projects. Given the importance of the subject, further research and studies are needed in this field to find more optimal methods for selecting and designing appropriate databases. This research can help develop deep learning models with higher accuracy and efficiency, and ultimately lead to significant advances in the field of artificial intelligence. It should be noted that the appropriate selection and design of the database is only one of the factors that affects the success of deep learning models. However, this factor, as one of the most important and fundamental aspects of any artificial intelligence project, must be carefully examined and evaluated.

2. Machine Learning and Deep Learning

In machine learning, a computer program is given a set of tasks to complete, and it is said that the machine has learned from its experience if its measured performance in these tasks improves over time as it obtains more and more practice completing them. This means that the machine is making judgments and forecasts based on historical data. Consider computer software that learns to diagnose cancer based on a patient's medical records. When it analyses medical investigation data from a larger population of patients, its performance will increase through the accumulation of knowledge (Sarker, 2021).

The number of cancer cases correctly predicted and detected, as verified by a seasoned oncologist, will serve as the performance metric. Machine learning is used in many different areas, including but not limited to robotics, virtual personal assistants (such as Google), video games, pattern recognition, natural language processing, data mining, traffic prediction, online transportation networks (such as Uber's surge pricing estimates), product recommendations, stock market forecasts, medical diagnoses, fraud predictions, agricultural advice, and search engine result refinement (such as Google's search engine) (Hassanien, Chang, & Mincong, 2021).

In artificial intelligence (AI), machine learning is the ability to automatically adapt with little to no human intervention, and deep learning is a subset of machine learning that uses neural networks to simulate the human brain's learning procedure. There is a wide gap between these two ideas. Although it needs more data to train on, deep learning can adapt to new circumstances and correct for its own faults. Contrarily, machine learning permits training on smaller datasets, but it requires more human intervention to learn and correct its errors. Machine learning relies on human intervention to categorise data and highlight attributes. In contrast, a deep learning system aims to acquire these qualities without any human input. In the simplest of explanations, machine learning functions like an obedient robot. Patterns in the data are analysed in order to make predictions. If you can imagine a robot that learns on its own, that is what deep learning is like. It can learn more intricate patterns and generate independent predictions.

Deep neural networks constitute a subfield of machine learning. It is a model of a network consisting of neurons with multiple parameters and layers between input and output. DL uses neural network topologies as its basis. Consequently, they are known as deep neural networks (Schmidhuber, 2015). DL provides autonomous learning of characteristics and their hierarchical representation at multiple levels. In contrast to conventional machine learning approaches (Table 01), this robustness is a result of deep learning's powerful process; in brief, deep learning's whole architecture is used for feature extraction and modification. The early layers do rudimentary processing of incoming data or learn simple features, and the output is sent to the upper layers, which are responsible for learning complicated features. Hence, deep learning is suited for handling larger data sets and greater complexity.

Table 1. Deep learning approaches.

	Advantages	Limitation
CNN	Highly effective for visual recognition	The quantity and ability of the training data have a significant impact on CNN performance.
	After learning a segment inside a certain area of an image, CNN can recognise that segment anyplace else in the picture.	extremely sensitive to noise
RNN	An RNN uses the same parameters throughout each phase, unlike a conventional neural network. This significantly lowers the number of parameters we need to memorise.	It is challenging for RNNs to monitor long-term dependence. This is particularly true when there are too many words between the noun and the verb in extended phrases and paragraphs.
	For unlabelled photos, RNNs may be used in conjunction with CNNs to produce precise descriptions.	RNNs can't be combined to create highly complex models. The reason for this is that the gradient decays over several layers as a result of the activation function employed in RNN models.
Generative Adversarial Networks (GANs)	GANs enable effective semi-supervised classifier training.	The effectiveness of the generator and discriminator is essential to GAN's success. Even if one of them fails, the entire system collapses.
	The produced data are practically indistinguishable from the original data due to the model's increased accuracy.	The discriminator and generator are distinct systems that were trained using various loss functions. It might thus take a long time to train the entire system.
Autoencoders	A model is produced that is mostly dependent on data rather than predetermined filters.	Sometimes training demands a lot of time.
	Very little complexity makes them simpler to train.	The information that emerges from the model may be hazy and confusing if the training data are not indicative of the testing data.
ResNets	In some situations, ResNets are more accurate and need fewer weights than LSTMs and RNNs.	If a ResNet has too many levels, faults may be difficult to see and difficult to transmit back fast and accurately. However, if the layers are too thin, the learning may not be as effective.
	A network may be built by adding tens of thousands of residual layers, which can then be trained.	

3. Types of Databases in Deep Learning

Databases in deep learning are divided into different types, each designed for specific applications. For example, image databases such as ImageNet and CIFAR-10 are widely used in training image recognition models (Hinton et al., 2006). These databases contain millions of images with different labels, which allow deep learning models to learn complex features. Also, text databases such as IMDB and Yelp are used for sentiment analysis and natural language processing (Deng, 2013).

Audio databases such as TIMIT and LibriSpeech are also used to train speech recognition and audio analysis models (Du & Swamy, 2019). These databases contain high-quality audio samples and corresponding labels that help models identify audio patterns. In addition, hybrid databases that contain multiple data types (image, text, audio) are also used for more complex applications such as multimedia systems and robotics (Vuong, 2021).

Also, custom databases that are designed for specific problems play an important role in the research and development of deep learning models. These databases usually contain specific data collected for specific applications such as medicine, agriculture, and finance (Benos et al., 2021). The design and use of custom databases can help improve the accuracy and efficiency of models.

4. Criteria for selecting a suitable database for deep learning

Choosing a suitable database for deep learning requires consideration of several criteria. The first criterion is data quality, which includes the accuracy, correctness, and completeness of the data (Han et al., 2012). High-quality data can help models learn more complex patterns and increase the accuracy of predictions. The second criterion is the size of the database. Larger databases usually allow models to be trained with more data and perform better on complex problems (Arel et al., 2010).

Data diversity is another important criterion. Highly diverse databases can make models more robust to changes and fluctuations in real data (Huang et al., 2020). Also, data access and privacy and ethical issues should be considered. Using databases that have been legally and ethically collected can avoid legal and social problems (Yuan et al., 2020).

The compatibility of the database with deep learning algorithms is also important. Some algorithms require specific data that must be considered when selecting a database (Gambella et al., 2021). Choosing the right database can significantly impact the performance and efficiency of deep learning models.

5. Data Preprocessing Methods to Improve the Performance of Deep Learning Models

Data preprocessing is a key step in preparing data for deep learning models. One common preprocessing method is data normalization, which helps models process data at a uniform scale (Kubat, 2017). This can improve the training speed and accuracy of models. In addition, data augmentation techniques are also used to increase data diversity and prevent overfitting (Karhunen et al., 2015).

Noise reduction methods are also used to improve data quality and increase model accuracy. These methods include filtering out irrelevant data and removing incorrect data (Ahmad et al., 2019). Feature selection techniques can also help reduce model complexity and improve its performance. These techniques extract important and relevant features from the data and prevent unnecessary data from entering the model (Janiesch et al., 2021).

Another important preprocessing method is data labeling. Labeled data helps models better identify patterns and relationships in the data (Srinivas et al., 2021). Using appropriate preprocessing techniques can significantly improve the performance of deep learning models and provide more accurate results.

6. Research Methodology

The main objective of this study is to investigate and analyze the factors affecting the selection and design of appropriate databases for deep learning models. The statistical population of this study includes experts and researchers active in the field of artificial intelligence and deep learning who have practical experience in this field. To select the sample, purposive sampling method was used to select people who have sufficient experience and knowledge in this field. The number of samples was determined based on the theoretical saturation criterion and 30 people were considered.

The data collection tool includes structured questionnaires and semi-structured interviews. The questionnaires are designed to collect comprehensive and accurate information about the experiences and views of experts in the field of database selection and design. Interviews are also conducted to collect qualitative and deeper data. Descriptive and inferential statistical methods are used to analyze data, and qualitative data is examined and analyzed using thematic analysis.

7. Research Findings

In this study, demographic information was collected from a sample of 30 experts and researchers active in the field of artificial intelligence and deep learning. Among the participants, 70% were male and 30% were female. The average age of the participants was 35 years, with an age range of 28 to 45 years. In terms of education level, 60% had a PhD, 30% had a Master's degree, and 10% had a Bachelor's degree. Also, 80% of the participants had more than 5 years of practical experience in the field of deep learning and database design. (Table 2)

In terms of geographical distribution, 50% of the participants worked in research and academic centers in large cities, while 30% worked in private companies and 20% in government institutions. Also, 40% of the participants participated in international projects, indicating their extensive interactions with the global scientific community.

Table 2: The effect of education level on the choice of database type

Education Level	Relational Database	NoSQL Database	Graph Database
Bachelor's Degree	40%	30%	30%
Master's Degree	30%	50%	20%
PhD	20%	60%	20%

This table shows that education level influences the choice of database type. People with higher education tend to use NoSQL databases more. This may be due to the complexity and need for more specialized knowledge in managing and using these types of databases. While people with a bachelor's degree tend to use relational databases, which have a simpler structure and management. These results indicate that education and education level can play an important role in the decision to choose more advanced technologies.

Table 3: The effect of practical experience on the use of different databases

Practical Experience (Years)	Relational Database	NoSQL Database	Graph Database
Less than 5 years	50%	30%	20%
5 to 10 years	30%	50%	20%
More than 10 years	20%	60%	20%

Practical experience also affects the choice of database type. As experience increases, there is a greater tendency to use NoSQL databases. This indicates that more experienced individuals are looking for technologies that offer greater scalability and flexibility (Table3). These results show that practical experience helps individuals tackle more complex challenges and use more advanced tools.

Table 4: Impact of Organization Type on Database Selection

Organization Type	Relational Database	NoSQL Database	Graph Database
Academic	40%	40%	20%
Private	30%	50%	20%
Government	50%	30%	20%

The type of organization also influences database selection (Table4). Private organizations are more inclined to use NoSQL databases, possibly due to the need for high flexibility and scalability in commercial projects. Government organizations, however, tend to prefer relational databases, which may be due to the need for structure and more security.

These results indicate that the work environment and organization type can influence technical decision-making. Private organizations might be seeking innovation and adopting newer technologies, whereas government organizations might prefer traditional technologies due to policies and regulations.

Table 5: Impact of International Projects on Database Selection

Participation in International Projects	Relational Database	NoSQL Database	Graph Database
Yes	30%	50%	20%
No	40%	40%	20%

Participation in international projects can also affect database selection (Table5). Individuals involved in international projects show a greater preference for NoSQL databases. This might be due to the need to coordinate and align with international teams and use common technologies. These results suggest that international interactions can facilitate the adoption of newer and more advanced technologies.

Table 6: Impact of Scalability Needs on Database Selection

Scalability Needs	Relational Database	NoSQL Database	Graph Database
Low	50%	30%	20%
Medium	30%	50%	20%
High	20%	60%	20%

Scalability needs are one of the key factors in database selection (Table6). For projects with high scalability needs, NoSQL databases are preferred due to their ability to scale horizontally and manage large volumes of data.

These results indicate that projects with high scalability needs seek technologies that allow for increased data volume and user numbers without sacrificing performance.

Table 7: Impact of Data Security Importance on Database Selection

Data Security Importance	Relational Database	NoSQL Database	Graph Database
Low	30%	40%	30%
Medium	40%	40%	20%
High	50%	30%	20%

Data security is also an important factor in database selection (Table 7). For projects where data security is of high importance, relational databases are more commonly used due to their strong security features and support for transactions and advanced access control.

These results show that projects with high security needs look for technologies that offer strong security features and support for transactions and advanced access control.

8. Discussion

In this study, the main objective is to investigate and analyze the factors affecting the selection and design of appropriate databases for deep learning models. Given the complexity and diversity of data needs in deep learning projects, the selection of an appropriate database can have a significant impact on the performance and efficiency of the models. In this regard, several factors such as education level, practical experience, type of organization, participation in international projects, scalability needs, and the importance of data security have been examined.

The results of this study show that the level of education has a direct impact on the selection of the type of database. People with higher education are more likely to use NoSQL databases, which require more specialized knowledge. These findings are consistent with the results of the study by Zou et al. (2022), which emphasizes the importance of machine learning in data management. They have shown that deeper expertise and knowledge can lead to the selection of more advanced technologies.

Practical experience is also another important factor in database selection. As experience increases, the tendency to use NoSQL databases increases. This suggests that more experienced individuals are looking for technologies that offer greater scalability and flexibility. These findings are consistent with Li and Zhou's (2022) study, which showed that practical experience can lead to the selection of more complex and efficient data management solutions.

The type of organization also influences the choice of database. Private organizations are more likely to use NoSQL databases, while government organizations prefer relational databases due to their need for greater structure and security. These results are consistent with Wang et al.'s (2023) study, which examined predictive and optimization models using deep learning. They showed that the organizational environment can influence technical decisions and that different types of organizations may have different needs, leading to the selection of different technologies.

The findings of this study emphasize that the selection and design of databases for deep learning models requires consideration of a set of individual and organizational factors that can lead to improved model performance and efficiency.

9. Conclusion

In this study, the influence of various factors on the selection and design of databases for deep learning models was investigated. The results showed that the level of education and practical experience play an important role in the selection of the type of database. People with higher education and more experience tend to use NoSQL databases due to the complexity and need for more specialized knowledge in managing and using these types of databases. Also, the type of organization and the scalability needs of the projects also have an impact on the selection of the database.

This study showed that the selection of the appropriate database for deep learning models requires attention to both individual and organizational factors. In particular, people who participate in international projects show a greater tendency to use NoSQL databases. This could be due to the need to coordinate and synchronize with international teams and use common technologies.

10. Recommendations

For deep learning projects, it is recommended that development teams pay special attention to training and increasing the knowledge of team members in the field of NoSQL databases so that they can take advantage of the scalability and flexibility capabilities of these technologies.

Organizations should conduct further research on the advantages and disadvantages of different databases, depending on the type of project and specific needs, and make their decisions based on detailed analysis and real project needs.

Developers should pay special attention to data security and, in projects where data security is of great importance, use databases that offer advanced security features and controls.

It is recommended that organizations and research teams exchange knowledge and experiences with international teams so that they can benefit from global experiences in selecting and designing databases.

11. Reference

1. Sami, Y., Richard, N., Gauchard, D., Estève, A., & Rossi, C. (2022). Selecting machine learning models to support the design of Al/CuO nanothermites. The Journal of Physical Chemistry A, 126(7), 1245-1254.
2. Yang, Z., & Gao, W. (2022). Applications of machine learning in alloy catalysts: rational selection and future development of descriptors. Advanced Science, 9(12), 2106043.
3. Zou, B., You, J., Wang, Q., Wen, X., & Jia, L. (2022). Survey on learnable databases: A machine learning perspective. Big Data Research, 27, 100304.
4. Li, G., & Zhou, X. (2022, May). Machine learning for data management: A system view. In 2022 IEEE 38th International Conference on Data Engineering (ICDE) (pp. 3198-3201). IEEE.
5. Saputra, N. A., Riza, L. S., Setiawan, A., & Hamidah, I. (2024). Choosing the appropriate deep learning method: A systematic review. Decision Analytics Journal, 100489.
6. Saleh, M., Jhanjhi, N., Abdullah, A., & Saher, R. (2022, February). Iotes (a machine learning model) design dependent encryption selection for iot devices. In 2022 24th International Conference on Advanced Communication Technology (ICACT) (pp. 239-246). IEEE.
7. Wang, S., Xia, P., Chen, K., Gong, F., Wang, H., Wang, Q., ... & Jin, W. (2023). Prediction and optimization model of sustainable concrete properties using machine learning, deep learning and swarm intelligence: A review. Journal of Building Engineering, 108065.
8. Lim, Y. G., Cho, Y. J., Sim, M. S., Kim, Y., Chae, C. B., & Valenzuela, R. A. (2020). Map-based millimeter-wave channel models: An overview, data for B5G evaluation and machine learning. IEEE Wireless Communications, 27(4), 54-62.
9. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput. Sci. 2021, 2, 160.
10. Hassanien, A.E.; Chang, K.C.; Mincong, T. (Eds.) Advanced Machine Learning Technologies and Applications; Springer Nature: Singapore, 2021; Volume 1141.
11. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. Neural Netw. 2015, 61, 85–117.
12. Arel, I., Rose, D.C., & Karnowski, T.P. (2010). Deep machine learning—A new frontier in artificial intelligence research [research frontier]. IEEE Comput. Intell. Mag., 5, 13–18.
13. Benos, L., Tagarakis, A.C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine Learning in Agriculture: A Comprehensive Updated Review. Sensors, 21, 3758.
14. Deng, L. (2013). Deep Learning: Methods and Applications. Found. Trends Signal Process., 7, 197–387.
15. Du, K.L., & Swamy, M.N. (2019). Neural Networks and Statistical Learning (2nd ed.). Springer Science & Business Media.
16. Gambella, C., Ghaddar, B., & Naoum-Sawaya, J. (2021). Optimization problems for machine learning: A survey. Eur. J. Oper. Res., 290, 807–828.
17. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann.
18. Hinton, G.E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. Neural Comput., 18, 1527–1554.
19. Huang, J., Chai, J., & Cho, S. (2020). Deep learning in finance and banking: A literature review and classification. Front. Bus. Res. China, 14, 13.
20. Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. Electron. Mark., 31, 685–695.
21. Karhunen, J., Raiko, T., & Cho, K. (2015). Unsupervised deep learning: A short review. In Advances in Independent Component Analysis and Learning Machines (pp. 125–142). Academic Press.
22. Kubat, M. (2017). An Introduction to Machine Learning. Springer International Publishing.
23. Srinivas, M., Sucharitha, G., & Matta, A. (2021). Machine Learning Algorithms and Applications. Wiley.
24. Vuong, Q. (2021). Machine Learning for Robotic Manipulation. Available online: https://arxiv.org/abs/2101.00755v1 (accessed on 11 April 2023).

25. Yuan, F.-G., Zargar, S.A., Chen, Q., & Wang, S. (2020). Machine learning for structural health monitoring: Challenges and opportunities. Sens. Smart Struct. Technol. Civ. Mech. Aerosp. Syst., 11379, 1137903.
26. Ahmad, J., Farman, H., & Jan, Z. (2019). Deep learning methods and applications. In Deep Learning: Convergence to Big Data Analytics (pp. 31–42). Springer.

CITATION
Reza Y., & Mostafa B. (2025). Selecting and designing appropriate databases for deep learning models. In Global Journal of Research in Engineering & Computer Sciences (Vol. 5, Number 2, pp. 104–111).
https://doi.org/10.5281/zenodo.15298240