Task B

Comparative Analysis of Methods in Supervised Learning and Deep Learning

Diego De Pablo

depablodiego@uma.es
Health Engineering. Málaga University.

This work explores supervised learning methodologies with a focus on deep learning (DL). We discuss the importance of splitting datasets into training and test sets, as well as cross-validation techniques for model evaluation. Further, we examine artificial neural networks (ANNs) and the advancements of DL over traditional neural networks, emphasizing innovations like convolutional layers and recurrent networks. Case studies of DL applications, such as AlphaFold in protein folding and DeepMind's AI in medical imaging, are also highlighted. Additionally, we address challenges such as overfitting and data limitations in biomedical fields. All this theoretically.

1 Introduction

In the rapidly evolving field of machine learning, supervised learning and deep learning have emerged as powerful solutions for complex problems across various domains. Supervised learning involves training models on labeled data to make predictions about unseen data, and it has been widely applied in tasks like classification and regression. Deep learning, a subset of machine learning, which focuses on using multi-layered artificial neural networks to process large volumes of data and automatically extract complex features. These networks, known as deep neural networks, are composed of several hidden layers between the input and output, allowing them to learn hierarchical representations of the data. [3]

Unlike traditional machine learning algorithms, where the relevant features are often designed manually, in deep learning deep neural networks are able to learn these features directly from raw data, such as images, audio or text.[3]

One of the critical challenges in developing machine learning models is establishing a fair comparison between different methods. This ensures that models are evaluated accurately, enabling practitioners to select the best model for their specific problem. To achieve this, it is crucial to carefully choose appropriate metrics from the literature and follow standardized procedures for training and testing.[6]

Through this study, it seek to gain a deeper understanding of supervised learning, explore critical aspects of model evaluation, and appreciate the transformative potential of deep learning in solving complex, real-world problems.

2 Supervised Learning Methodology

The report will address a number of key questions related to supervised methods and deep learning (DL). Starting with an introduction to supervised learning methodology, it will discuss the importance of splitting data into training and test sets to avoid generalization and model overfitting issues. Then, it will explore the concept of cross-validation, a technique that improves model performance evaluation and prevents overfitting by using different subsets of the data for training and validation.

Next, it will delve into artificial neural networks, a fundamental component of machine learning, and how deep neural networks (DL) go beyond traditional forward propagation networks. It will discuss what distinguishes deep learning models, exploring their advanced capabilities compared to classical neural networks.

In addition, it will present some notable cases where companies such as Google have achieved surprising results using DL in different domains, such as protein structure prediction or medical image analysis. A

crucial challenge in training DL models, overfitting, will also be discussed, and how deep networks employ techniques such as regularization and dropout to mitigate this problem.

Finally, the limitations that lack of data presents in the application of DL models, especially in the biomedical field, will be examined, and available techniques, such as transfer learning and synthetic data generation, will be reviewed to overcome this barrier and improve the applicability of these models in contexts where data is scarce.

2.1 About Supervised Learning Methodology

Supervised learning is a type of machine learning where a model is trained using labeled data, meaning that each training example is paired with the correct output. The goal of supervised learning is to learn a mapping from inputs to outputs so that the model can make accurate predictions when given new, unseen data.[4]

For example if you have a dataset of medical records where each record includes features like age, blood pressure, and cholesterol levels (inputs) along with whether or not the patient has a disease (output), a supervised learning algorithm could be trained on this data to predict whether a new patient is at risk of the disease.

Key Concepts in Supervised Learning: *Why do we need to split the dataset into training and test sets?*

When developing a supervised learning model, it is crucial to divide the dataset into training and test sets. The training set is used to train the model, meaning the model learns from this subset of data by adjusting its parameters to minimize errors in predictions. The test set, on the other hand, is used to evaluate the model's performance on data it has never seen before. This separation ensures the model's ability to generalize to new, unseen data, which is essential for robust performance in real-world scenarios.

You can see the figure 1 where you can see how a dataset is normally treated when applying a supervised method.

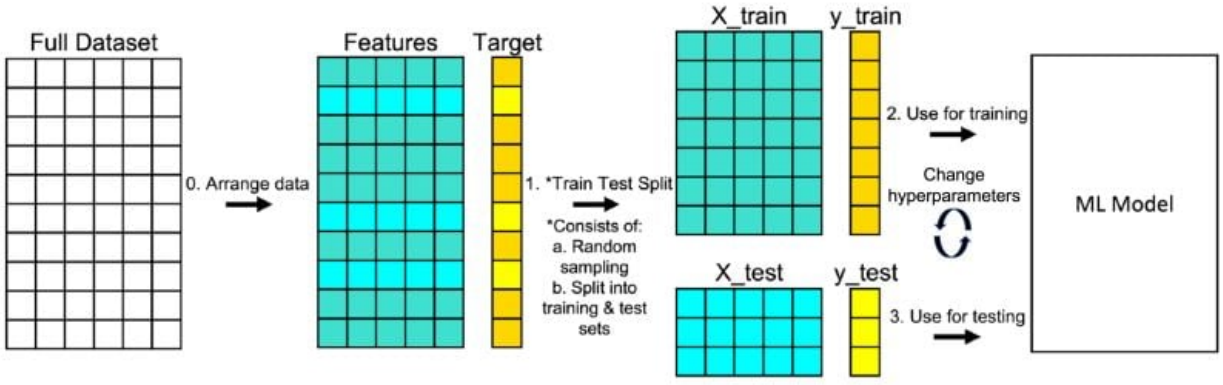

Fig. 1. Train test split procedure. — Image: Michael Galarnyk[5]

The train-test split is a model validation procedure that helps simulate how a model would perform on new, unseen data. The following steps describe how the procedure works:

1. **Arrange the Data:** Ensure that the data is in the correct format.For example in the case of python with scikit-learn, this means separating the dataset into "Features" (inputs) and "Target" (output).
2. **Split the Data:** The dataset is split into two parts: 70% (or another percentage depend of your data) is used for training, and 30% is used for testing. The image shows how the data is divided into "X_train," "X_test," "y_train," and "y_test."

3. **Train the Model:** The model is trained on the training data ("X_train" and "y_train") to learn from the input-output relationships.
4. **Test the Model:** The trained model is tested using the test set ("X_test" and "y_test") to evaluate its performance.

As mentioned, dividing the data for training and testing produces very good results and also allows for a more honest and robust evaluation of the model. This is why there are more methods for dividing initial data, such as cross-validation.

Cross-validation *What is cross-validation?*

Is a powerful technique used to evaluate the performance of a supervised learning model more reliably than a simple train-test split. Instead of splitting the dataset into just one training set and one test set, **k-fold cross-validation** involves dividing the dataset into k equally sized subsets or folds (see figure 2).

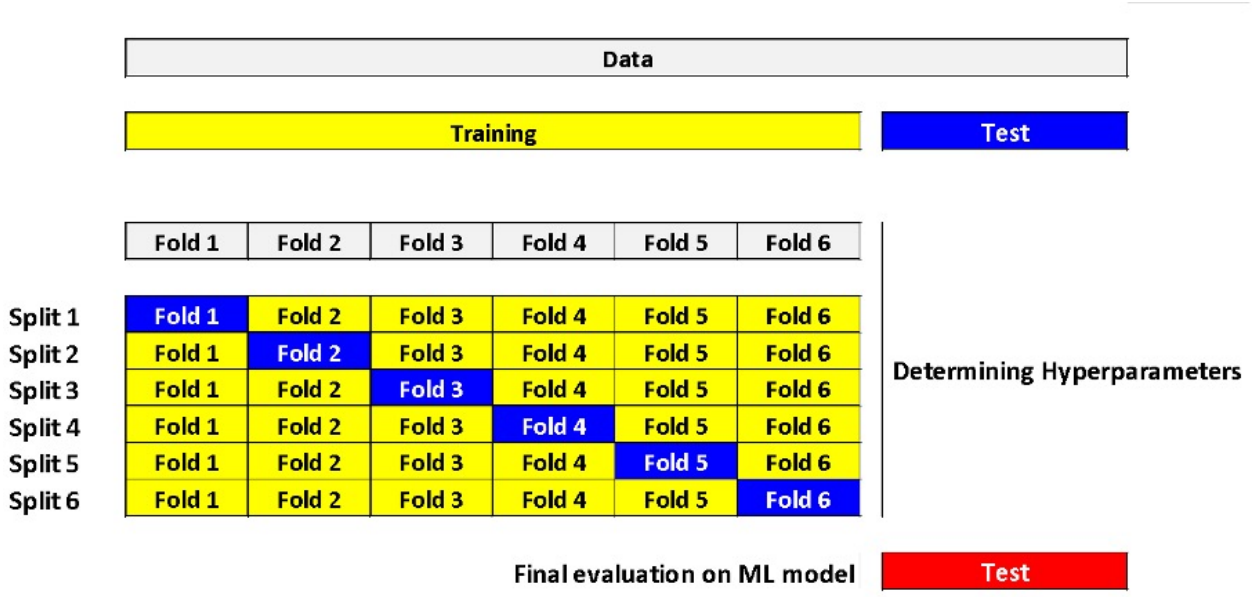

Fig. 2. The k-fold cross-validation randomly splits the original dataset into k number of folds[8]

The process works as follows:

1. The model is trained on $k-1$ of the folds and tested on the remaining fold.
2. This process is repeated k times, with each fold used as the test set once.
3. The final performance metric is averaged over all k trials, giving a more robust estimate of the model's ability to generalize.

In practice, common variations of cross-validation include:

- **k-fold cross-validation:** Typically with $k = 5$ or $k = 10$, this method divides the dataset into k folds and tests the model on each fold.
- **Leave-one-out cross-validation (LOOCV):** Where each data point acts as a single test set.

Cross-validation ensures that the model's evaluation is more stable, reliable, and less dependent on any specific train-test split configuration.[8]

3 Neural Networks and Deep Learning

Another artificial intelligence algorithm we will introduce the fundamental concepts is artificial neural networks (ANNs), followed by a deeper dive into deep learning (DL), which is an extension of ANNs.

Artificial Neural Networks *What are Artificial Neural Networks (ANNs)?*

Artificial Neural Networks (ANNs) are computational models inspired by the human brain (see the figure 3). They consist of interconnected nodes (also called neurons) organized in layers: an input layer, one or more hidden layers, and an output layer. Each node in a layer receives input from the previous layer, processes it, and passes the result to the next layer. The strength of the connection between nodes is determined by weights, which are learned during training.[1]

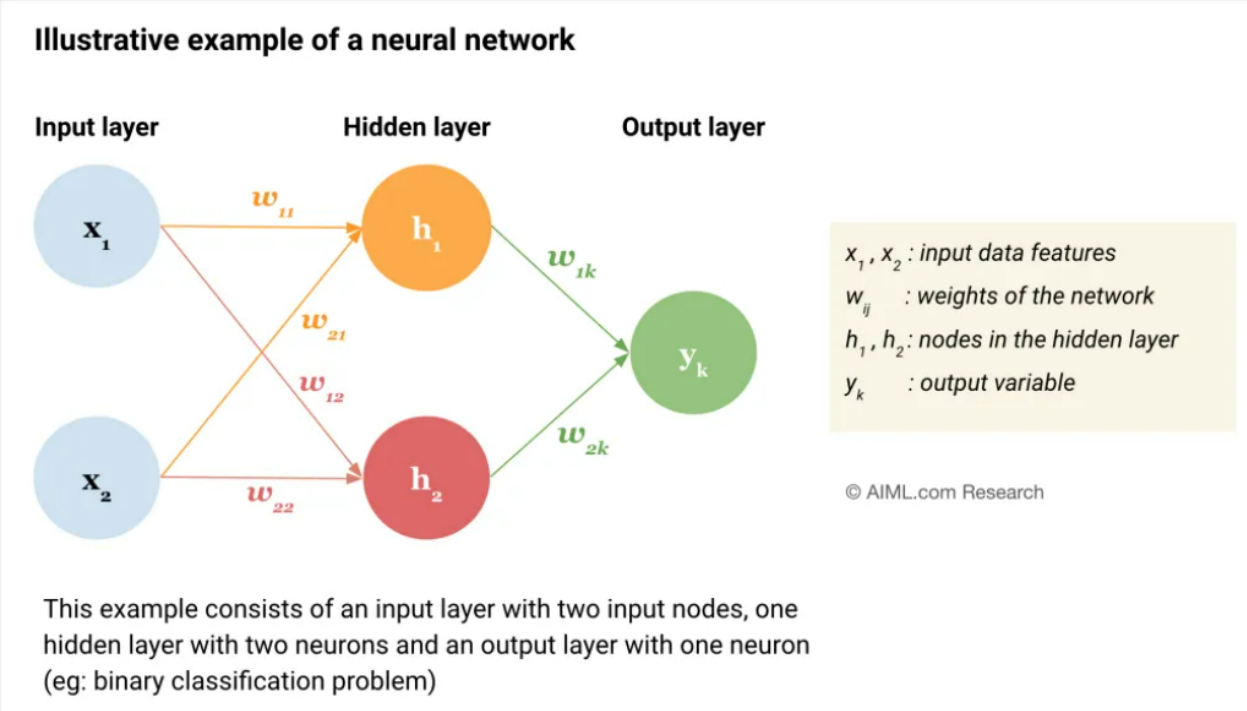

Fig. 3. A simple neural network, an illustrative example[1]

- **Input Layer:** The input layer consists of neurons that receive the raw data and pass it to the hidden layers for further processing.
- **Hidden Layers:** These layers process the input data using activation functions to introduce non-linearity, allowing the model to learn complex patterns.
- **Output Layer:** The output layer produces the final result of the network, whether it's a classification label, continuous value, or probability.

ANNs are widely used in machine learning tasks such as classification, regression, and pattern recognition.[1]

Deep Learning *What is Deep Learning (DL)?*

Is a subset of machine learning that builds upon artificial neural networks by introducing **deep architectures**. These architectures consist of multiple hidden layers, enabling models to learn complex and

hierarchical representations of data. With deep learning, neural networks can automatically learn high-level features from raw data without the need for manual feature extraction. This makes DL particularly powerful in fields like computer vision, speech recognition, and natural language processing.[9]

The deeper the network (more layers), the more complex patterns it can capture, making deep learning models more capable of handling high-dimensional data and tasks that require advanced feature learning.

Deep Learning against Neuronal Network *What is New in DL Models Compared to Traditional Feed-forward Neural Networks?*

Several innovations have contributed to the success of deep learning models compared to traditional feedforward neural networks:[9]

- **Convolutional Layers:** Used primarily in Convolutional Neural Networks (CNNs), these layers are designed to automatically detect spatial hierarchies in data, such as edges in images, making them highly effective for image classification and recognition tasks.
- **Recurrent Neural Networks (RNNs):** These networks include feedback loops that allow information to persist, making them ideal for sequence data like time series or text, where past inputs influence future predictions.
- **Backpropagation:** A key advancement in neural networks, backpropagation allows for efficient training by adjusting weights based on the error in the output, making it possible to train deeper networks and improve their performance.
- **Other Advancements:** Techniques such as dropout, batch normalization, and the use of GPUs for faster computation have significantly improved the training and performance of deep learning models.

These innovations make deep learning models more powerful and flexible compared to traditional feed-forward neural networks, allowing them to solve more complex problems and generalize better to unseen data.[9]

4 Applications of Deep Learning

Deep learning has revolutionized many fields, demonstrating transformative results in various domains. Below are two notable case studies where deep learning has produced astonishing results:

4.1 AlphaFold and Protein Structure Prediction

AlphaFold, developed by DeepMind, has made significant advances in predicting protein structures. This has been a longstanding challenge in biology, and AlphaFold's ability to predict the 3D structure of proteins based on their amino acid sequences has profound implications for drug discovery and disease research [2] (See Figure 4).

In the Figure 4, panel a shows AlphaFold's performance on the CASP14 dataset (n = 87 protein domains) compared to the top 15 entries out of 146, with group numbers assigned by CASP. The data represent the median and 95% confidence interval, estimated from 10,000 bootstrap samples. Panel b displays our prediction of CASP14 target T1049 (PDB 6Y4F, blue) alongside the true structure (green), excluding four C-terminal residues as B-factor outliers. Panel c shows the accurate prediction of a zinc-binding site in target T1056 (PDB 6YJ1). Panel d illustrates the correct domain packing of the 2,180-residue chain in CASP target T1044 (PDB 6VR4), predicted using AlphaFold without intervention. Panel e depicts the model architecture, with arrows indicating information flow and array shapes specified for sequences (s), residues (r), and channels (c).[10]

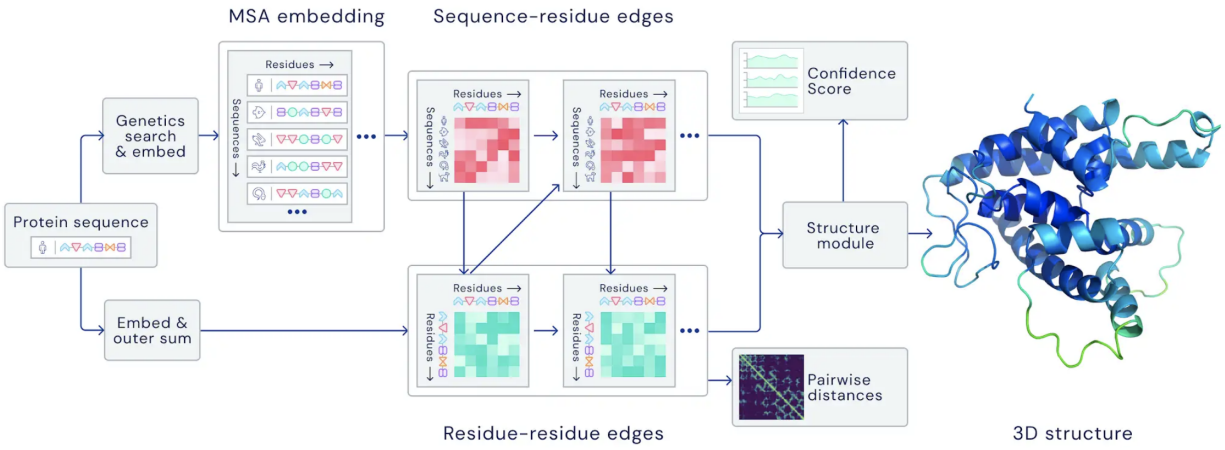

Fig. 4. Example of AlphaFold[10]

4.2 Google's DeepMind AI in Medical Imaging

DeepMind's AI has achieved remarkable results in medical imaging, particularly in diagnosing diseases such as diabetic retinopathy and age-related macular degeneration (AMD). By analyzing retinal scans, the AI system matches or exceeds the diagnostic accuracy of expert doctors, aiding in early detection and improving patient outcomes.

In the figure 5 Training process for a deferral AI model in medical imaging. Medical cases are processed by a **Predictive AI Model**, producing *predictive confidence scores*. In parallel, a **Clinical Workflow** provides *retrospective clinician opinions*. These inputs are used to train the **Deferral AI Model**, which leverages *biopsy-proven ground truth labels* to determine when to defer to human experts based on confidence levels and clinician feedback.

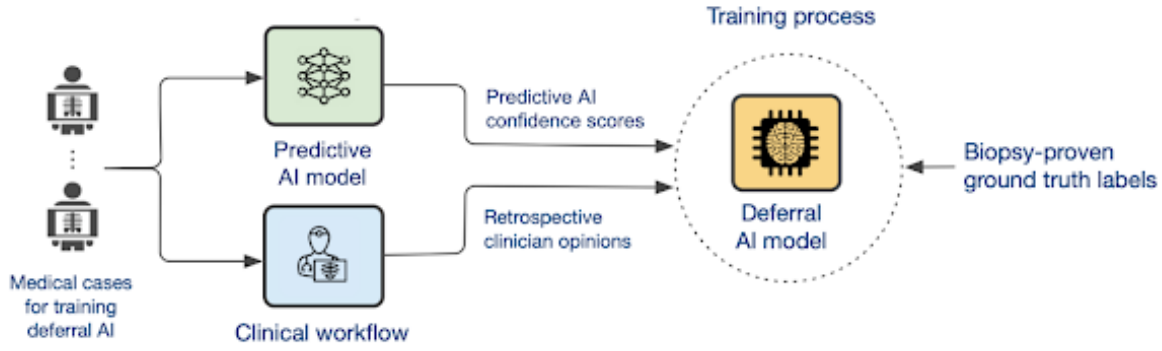

Fig. 5. Example of AlphaFold[10]

5 Challenges in Deep Learning

Deep learning models face a variety of challenges, two of the most prominent being overfitting and data limitations, especially in specialized fields such as biomedical applications.

5.1 Overfitting and Mitigation Techniques

Overfitting is one of the shortcomings in machine learning that hinders the accuracy and performance of the model. occurs when a model learns to perform exceptionally well on the training data, but fails to generalize to unseen data. This happens when the model becomes too tailored to the specific training set, capturing noise and patterns that don't apply to other data. The result is reduced accuracy and performance on real-world data, which undermines the model's effectiveness in practical applications.[7]

- **Dropout:** A regularization method that randomly disables a fraction of neurons during training, forcing the network to learn more robust features.
- **Regularization:** Techniques like L1 or L2 regularization add penalty terms to the loss function, discouraging the model from fitting excessively to the training data.
- **Early Stopping:** This technique halts training when the model's performance on a validation set starts to degrade, preventing the model from overfitting to the training data.

5.2 Addressing Data Limitations in Biomedical Applications

In biomedical applications, deep learning models often struggle with limited data. This is especially challenging in fields like medical imaging or genomics, where collecting large, labeled datasets can be costly and time-consuming. Several techniques can help alleviate this issue:

- **Data Augmentation:** This involves creating modified versions of the training data (e.g., rotating, zooming, or flipping images) to increase the dataset's size and diversity, helping the model generalize better.
- **Transfer Learning:** Pretrained models on large datasets are fine-tuned on smaller biomedical datasets, leveraging knowledge learned from a broader task to improve performance on a specific biomedical task.
- **Synthetic Data Generation:** Techniques such as generative adversarial networks (GANs) can be used to generate synthetic data that mimics real-world biomedical data, further expanding the available dataset for training deep learning models.

6 Conclusion

Validating machine learning models is essential to ensure their accuracy, honesty, and generalizability. A reliable model is one that not only performs well on training data but also generalizes effectively to new, unseen data. Techniques like cross-validation and careful data splitting help in assessing a model's robustness, enabling us to avoid pitfalls like overfitting.

When working with deep learning (DL), while the technology has proven transformative in fields such as healthcare and protein folding, it is far from perfect. The complexity of DL models makes it essential for developers to understand the underlying processes. This knowledge ensures that they can apply appropriate techniques, such as dropout or data augmentation, to avoid overfitting and manage data limitations effectively. Ultimately, the success of DL models relies on a careful balance between model complexity, data quality, and proper validation.

7 Repository Github

Further information, including the source code and full project documentation, can be accessed through the GitHub repository. Click here to go to the repository.

References

1. Karphatij Andrej. What is the basic architecture of an artificial neural network (ann)?, 2024. Accessed: 2024-11-10.
2. EMBL-EBI. Alphafold, 2024. Accessed: 2024-11-10.
3. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. Comprehensive introduction to deep learning, covering theory and practice.
4. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009. A foundational text on machine learning and statistical modeling, covering supervised learning and model evaluation techniques such as cross-validation.
5. Built In. What is the train-test split and why is it important?, 2024. Accessed: 2024-11-10.
6. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. Seminal paper that provides an overview of the field of deep learning, including key advances and applications.
7. Protección Datos LOPD. ¿qué es el overfitting en el aprendizaje automático?, 2024. Accessed: 2024-11-10.
8. Kili Technology. Cross-validation in machine learning, 2024. Accessed: 2024-11-10.
9. Turing. Deep learning vs machine learning: The ultimate battle, 2024. Accessed: 2024-11-10.
10. Papers with Code. Alphafold, 2024. Accessed: 2024-11-10.