Lam, C. (2011). *Hadoop in Action*. Manning Publications Co.: Stamford, CT. (Lam, 2011).

Hadoop in Action" is a comprehensive guide that introduces readers to Hadoop and its ecosystem. The book provides a solid foundation in understanding the basics of Hadoop, including its architecture, file system (HDFS), and core components like MapReduce. It demonstrates how Hadoop can be leveraged to process large datasets efficiently. The book includes case studies that showcase Hadoop in real-world applications:

Converting 11 million image documents from the New York Times archive.

Mining data at China Mobile.

Recommending the best websites at StumbleUpon.

Building analytics for enterprise search in IBM’s Project ES2.

Furthermore, the book covers advanced topics such as setting up a Hadoop cluster, data processing with Hive and Pig, and introduces other components of the Hadoop ecosystem like HBase, ZooKeeper, and Sqoop. This book equips readers with the necessary knowledge to get started with Hadoop and understand its significance in the realm of Big Data. This source helps to justify the importance of Hadoop in Big Data analytics and offers a thorough explanation of what a distributed file system is and how hadoop entered in the scene by providing computational capabilities over large amounts of data.

The Past, Present, and Future of Machine Learning APIs (Atakan Cetinsoy et al., 2016).

This paper discusses the evolution of ML APIs, tracing their journey from the era when researchers relied on pen and paper, to the present, where IT systems deploy these ML models. It also addresses current challenges, emphasizing the treatment of data before implementing an ML model. The authors highlight solutions such as Machine Learning as a Service (e.g., Azure ML, AWS) and REST APIs (e.g., BigML), showing how these technologies are valuable because they provide evidence of large companies creating and using APIs for modelling their ML models.

Deep Machine Learning and Neural Networks: An Overview (Mishra and Gupta, 2017).

According to Mishra and Gupta (2017), deep learning surpasses traditional ML models in its ability to perceive text and images. Neural Networks, as a crucial component of deep learning, are discussed in depth—specifically, ANNs and CNNs, which are key to this paper. It concludes that NNs are among the more popular techniques for solving deep learning problems.

Introduction to convolutional neural network using Keras; an understanding from a statistician. (Lee and Song, 2019).

In this study, Lee and Song (2019) focus on examining parameter estimation procedures for deep neural networks, as well as the structures of CNN models, ranging from basic to advanced techniques. The authors are also determined to demonstrate the critical steps in CNNs that enhance image classification performance on the CIFAR-10 dataset using Keras. Their conclusions suggest that utilizing multiple stacks of convolutional layers along with batch normalization can lead to improved predictions. Another important finding that is relevant to this paper is that Keras, due to its popularity as a neural network API, enables individuals to quickly familiarize themselves with deep learning methodologies.

Accelerating Relational Databases by Leveraging RemoteMemory and RDMA (Li et al., 2016).

Li et al., (2016) studied the crucial role of memory in RDBMS, especially comparing physical memories (SSD or HDD) with cloud memory (remote direct access, RDMA). This paper examines four scenarios to demonstrate that remote memory hosted in servers outperforms traditional physical memory systems. It is relevant to note that the state of the art of RDBMS has dramatically improved thanks to cloud-based solutions.

A Unified Metamodel for NoSQL and Relational Databases. (Candel, Sevilla Ruiz and García-Molina, 2022).

Candel, Sevilla Ruiz, and Garcia-Molina (2022) are determined to prove that a unified metamodel for NoSQL and relational databases represents the future in the field of databases, especially as NoSQL technologies have gained popularity recently. In their conclusion, the authors present a metamodel named U-Schema, where both NoSQL and RDBMS are optimized. Pertinent to this paper is an understanding of the current state of the art concerning RDBMS.

A survey on RDBMS and NoSQL Databases MySQL vs MongoDB (Palanisamy and Suvitha Vani, 2020).

Palanisamy and Suvitha Vani (2020) conducted a survey comparing the concepts of NoSQL and RDBMS, including their limitations, and also addressed the advantages and types of NoSQL databases. Their research concluded that a NoSQL database is an excellent choice when the data is structured, and the volume is not large, while NoSQL is preferable for unstructured data or for structured data with the potential for rapid growth. Essential to this paper is the way this article summarizes the points where NoSQL overtakes RDBMS.

Evolution of Hadoop and Big Data Trends in Smart World (Neeta Awasthy and Nikhila Valivarthi, 2023).

Neeta Awasthy and Nikhila Valivarthi (2023) focus their study on how large corporations tackle the challenges of Big Data by implementing HDFS. The index building for Google Search involves massive datasets and performs statistical analysis in indexing through large-scale batch processing. Facebook manages two primary clusters that store approximately 12 terabytes of data each. Additionally, eBay maintains substantial clusters designed to manage the data of 180 billion active users. To this paper, the value of this study lies in how it demonstrates the real-world applications of HDFS.

The Unreasonable Effectiveness of Data (Halevy, Norvig, & Pereira, 2009).

Halevy, Norvig, and Pereira (2009) argue that large volumes of data can be more valuable than complex algorithms in developing artificial intelligence systems. They maintain that large-scale data can improve language processing and that simple algorithms can outperform complex ones when coupled with massive datasets. How this study fits into the paper is significant because it demonstrates an early stage of what is now known as an API; the authors used web-based data to train their models.

AutoKeras: An AutoML Library for Deep Learning (Jin et al., 2023).

Jin, Chollet, Song, and Hu (2023) present AutoKeras, an Automated Machine Learning (AutoML) library designed to simplify the application of deep learning. It offers a solution for challenges in model selection and hyperparameter tuning, thereby making deep learning accessible to those with limited technical expertise. The library provides a user-friendly interface, assisting inexperienced users in addressing machine learning problems with minimal coding required. This study represents a significant step towards democratizing deep learning technology and aligns with the promising future of ML APIs discussed in this paper.

Deep Learning (LeCun, Bengio, & Hinton, 2015).

LeCun, Bengio, and Hinton (2015) review the filed of deep learning emphasizing its ability to create computational models that recognize complex patterns through multiple processing layers. The authors describe how deep learning has improved areas like speech recognition, visual recognition, object detection, drug discovery, and genomics. All these advancements are thanks to the use of backpropagation algorithms. It also mentions that deep convolutional networks are great for image, video, audio processing, text processing, and speech processing.

ImageNet Classification with Deep Convolutional Neural Networks (Krizhevsky, Sutskever, & Hinton, 2012).

Krizhevsky, Sutskever, and Hinton (2012) explain the development of a deep convolutional neural network (CNN) that outperforms prior models in classifying high-resolution images in the ImageNet challenge. The CNN features 60 million parameters and 650,000 neurons across five convolutional and three fully connected layers. This study is one of the first to introduce the term 'dropout' as a method to prevent overfitting

Attention Is All You Need (Vaswani et al., 2017).

Vaswani et al. (2017) introduce a novel neural network architecture known as the Transformer, which represents an evolution from recurrent and convolutional neural networks (RCNN). The Transformer employs attention mechanisms, enabling better parallelization, reducing training times, and enhancing performance on machine learning translation tasks. This paper concludes that the Transformer is a superior model to its predecessors in terms of translation models.

An Evaluation of Training Size Impact on Validation Accuracy for Optimized Convolutional Neural Networks (Barry-Straume et al., 2019).

Barry-Straume, Tschannen, Engels, and Fine (2018) assess how varying training set sizes influence the validation accuracy of CNNs. The study determines the optimal data volume required to achieve maximum accuracy during model validation. Furthermore, the research indicates that larger datasets can significantly enhance the predictive capabilities of CNNs. This study is crucial to the present paper as it introduces the concept of validation. In the implementation phase, validation was conducted to assess overfitting. Initially, the neural network and the specific CNN in focus here were prone to overfitting; however, through validation, this was effectively mitigated.

Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis (Alwosheel, van Cranenburgh and Chorus, 2018).

Alwosheel, van Cranenburgh, and Chorus (2018) examine the appropriate sample size for ANNs in discrete choice modelling. They fill a gap in empirical guidelines by establishing a rule-of-thumb based on Monte Carlo analyses of both synthetic and real data. Their research suggests that a dataset size fifty times the number of weights in the ANN is more effective than the traditionally used benchmark of a dataset only ten times the weight count. This research illuminates a section of this paper, where, in phase 2, an overfitted NN was corrected by increasing the dataset size, thus resolving the overfitting issue.

Deep Learning with Python. (Chollet, 2018, pp. 253-259).

Chollet (2018) provides great examples of best practices for deep learning, particularly the use of Keras as an API for modelling CNNs. Chollet demonstrates how Keras streamlines data processing in a CNN structure, from processing IMDB reviews to constructing network layers. This example was instrumental to this paper, showcasing Keras robustness and ease of use in the context of neural networks.

Regularization of deep neural networks with spectral dropout (Khan, Hayat and Porikli, 2019).

Khan, Hayat, and Porikli (2019) introduce 'Spectral Dropout,' an enhancement of the dropout technique newly introduced for the ImageNet challenge in 2012 by Krizhevsky, Sutskever, and Hinton (2012). Spectral Dropout augments traditional CNNs with a decorrelation transform and tackles overfitting by mitigating weak and noisy Fourier domain coefficients of network activations. The research concludes that this method's efficacy surpasses current regularization methods and increases network training speed. This paper is crucial in addressing overfitting in CNN phase 1, highlighting that since 2012, researchers have been using and improving dropout, leading to its implementation in CNN phase 2.

NoSQL: The Future of Big Data Analytics and Comparison with RDBMS (Arshad et al., 2023).

This article compares NoSQL technologies with traditional RDBMS in the context of Big Data analytics. The paper describes NoSQL as "Not Only SQL" and categorizes these databases into key-value stores, document databases, wide-column stores, and graph databases, with Hadoop classified as a wide-column store. It outlines the evolution of Big Data from megabytes and gigabytes to terabytes and petabytes, constantly challenging the industry to develop new storage solutions to meet escalating demands. The nature of Big Data is also elucidated in terms of its volume, variety, velocity, and variability.

Furthermore, the paper details the ACID properties of RDBMS—atomicity, consistency, isolation, and durability—and compares them with the CAP theorem of NoSQL, which emphasizes strong consistency, high availability, and partition tolerance. The authors conducted a survey among relevant IT companies, revealing a preference for NoSQL technologies when managing Big Data. The study concludes that applications dealing with Big Data tend to perform better in NoSQL environments.

***How to mention the sources in the paper:***

In a literature review, it's standard to mention sources through in-text citations rather than writing down the full name of the paper each time you refer to it. The Harvard referencing style, for example, would have you include the author’s surname and the year of publication in parentheses within the text. This method allows your narrative to flow more smoothly while still giving credit to the original authors and enabling readers to find the detailed reference in your bibliography or reference list.

Here’s a brief example to illustrate:

Incorrect: "In the literature review, writing down the name of the paper each time for reference might disrupt the flow of reading."

Correct: "According to Smith (2020), integrating in-text citations within the narrative of a literature review enhances readability and allows for a smoother flow of information."

Remember, the exact format for your in-text citations and reference list entries will depend on the specific guidelines provided by your instructor or the preferred citation style of your academic discipline.

**Paper References:**

Alwosheel, A., van Cranenburgh, S. and Chorus, C.G., 2018. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. Journal of Choice Modelling. https://doi.org/10.1016/j.jocm.2018.07.002

Arshad, M., Brohi, M.N., Soomro, T.R., Ghazal, T.M., Alzoubi, H.M. and Alshurideh, M., 2023. NoSQL: Future of Big Data Analytics Characteristics and Comparison with RDBMS. In: *The Effect of Information Technology on Business and Marketing Intelligence Systems*, pp.1927-1951. doi: 10.1007/978-3-031-12382-5\_106.

Barry-Straume, J., Tschannen, A., Engels, D.W. and Fine, E., 2018. An Evaluation of Training Size Impact on Validation Accuracy for Optimized Convolutional Neural Networks. *SMU Data Science Review*, 1(4), Article 12. https://scholar.smu.edu/datasciencereview/vol1/iss4/12

Candel, C.J.F., Sevilla Ruiz, D. and García-Molina, J.J., 2022. A unified metamodel for NoSQL and relational databases. *Information Systems, 104*, p.101898. <https://doi.org/10.1016/j.is.2021.101898>.

Cetinsoy, A., Martin, F.J., Ortega, J.A. and Petersen, P., 2016. The Past, Present, and Future of Machine Learning APIs. In*: PAPIs 2015 - Proceedings of the 2015 Conference on Predictive APIs and Apps*, vol. 50, JMLR: Workshop and Conference Proceedings, pp.43-49.‌

Chollet, F. (2018). *Deep Learning with Python*. Shelter Island, NY: Manning Publications. pp. 253-259.

Halevy, A., Norvig, P., & Pereira, F. (2009). *The Unreasonable Effectiveness of Data*. IEEE Intelligent Systems, 24(2), 8-12.

Jin, H., Chollet, F., Song, Q. and Hu, X. (2023). AutoKeras: An AutoML Library for Deep Learning. Journal of Machine Learning Research, 24, 1-6.

Khan, S.H., Hayat, M. and Porikli, F., 2018. Regularization of deep neural networks with spectral dropout. *Neural Networks.* https://doi.org/10.1016/j.neunet.2018.09.009

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet classification with deep convolutional neural networks.* Advances in Neural Information Processing Systems, 25.

Lam, C. (2011). *Hadoop in Action*. Manning Publications Co.: Stamford, CT.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning. Nature,* 521(7553), 436-444.

Lee, H. & Song, J. (2019). Introduction to convolutional neural network using Keras: An understanding from a statistician. *Communications for Statistical Applications and Methods*, 26(6), 591–610. https://doi.org/10.29220/csam.2019.26.6.591

Li, F., Das, S., Syamala, M. and Narasayya, V.R., 2016. Accelerating Relational Databases by Leveraging Remote Memory and RDMA. In: *Proceedings of the 2016 International Conference on Management of Data - SIGMOD ’16.* https://doi.org/10.1145/2882903.2882949.

Mishra, C. and Gupta, D.L., 2017. Deep Machine Learning and Neural Networks: An Overview. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 6(2), pp.66-73. <https://doi.org/10.11591/ijai.v6.i2.pp66-73>.

Neeta Awasthy and Nikhila Valivarthi (2023). Evolution of Hadoop and Big Data Trends in Smart World. In S. Awasthi, G. Sanyal, C.M. Travieso-Gonzalez, P. Kumar Srivastava, D.K. Singh & R. Kant (Eds.), *Sustainable Computing* (pp.148-186). Cham: Springer. https://doi.org/10.1007/978-3-031-13577-4\_6

Palanisamy, S. and SuvithaVani, P., 2020. A survey on RDBMS and NoSQL Databases: MySQL vs MongoDB. In: *2020 International Conference on Computer Communication and Informatics (ICCCI -2020)*, 22-24 January 2020, Coimbatore, India. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need.* Advances in Neural Information Processing Systems, 30.