DOI: https://doi.org/10.3126/nccj.v9i1.72262

Original Research Article

Role of Statistics in Artificial Intelligence Technology

Dila Ram Bhandari[1], Michael Baron[2], Kapil Shah[3], & Sharda Kandel[4]

[1] Principal Author
Lecturer, Nepal Commerce Campus, Tribhuvan University
Email: dilabhandarig@gmail.com

[2] Corresponding Author
Professor, American University,
Email: baron@american.edu

[3] Co-Author
Lecturer, Nepal Commerce Campus, Tribhuvan University
Email: kapil.shah@ncc.tu.edu.np

[3] Co-Author
Chairperson, PULSE Nepal Research Centre Pvt. Ltd.
Email: ksharda2071@gmail.com

Keywords
Artificial intelligence, Data science, Machine learning, Statistics,

JEL Classification Codes:
C18, C55, O33

Online Access

DOI:
https://doi.org/10.3126/nccj.v9i1.72262

How to Cite APA Style
Bhandari, D. R., Baron, M., Shah, K., & Kandel, S. (2024). Role of Statistics in Artificial Intelligence Technology. *NCC Journal, 9*(1), 133-139

Abstract
Artificial intelligence (AI) research and applications have sparked a broad scientific, economic, social, and political debate. Statistical approaches and techniques are fundamental to AI because they allow robots to learn from data and make intelligent decisions. It's possible even to view statistics as a fundamental component of AI. Statistics is an ideal partner for other disciplines in teaching, research, and practice because of its specialized knowledge of data evaluation, which begins with the correct phrasing of the research question and continues through a study design stage to analysis and interpretation of the results. This work aims to demonstrate how statistical methodology is relevant to the development of AI. In terms of methodological development, planning, research design, evaluation of data quality and collection, distinction of causation and relationships, and evaluation of result uncertainty, we talk about the contributions of statistics to the field of artificial intelligence. This study thoroughly analyzes the crucial role statistics play in AI, demonstrating the numerous applications of statistical ideas like probability, regression, classification, and clustering. The essay highlights the value of statistical analysis in handling uncertainty, making predictions, and training AI models all of which improve the efficiency and precision of AI systems.

Introduction
Artificial Intelligence has rapidly evolved, integrating into various facets of modern life and reshaping industries with its radical competencies. The phrase "artificial intelligence" refers to the ability of machines to simulate cognitive processes including perception, learning, reasoning, problem-solving, and decision-making. At the core of AI's transformative power is a robust reliance on statistical methods and principles. Statistics play a critical function in AI by aiding systems to acquire data, make informed decisions, and predict outcomes with a degree of certainty.

Statistical methods are foundational to various key AI technologies. Machine learning, a subfield of AI, vitally depends on statistical methods for model training and validation. Algorithms such as linear regression, decision trees, and neural networks rely on statistical principles to interpret data and derive predictive models. For instance, in supervised learning, statistical

methods help in estimating the relationships between variables, optimizing model functioning, and validating the precision of predictions (Bishop, 2006; Hastie et al., 2009).

Bayesian statistics, a branch of statistics engrossed in probability inference, plays a significant role in AI. Bayesian methods are used to update the probability of a hypothesis as more evidence or data becomes available, which is crucial for tasks such as decision-making under uncertainty and probabilistic reasoning (Gelman et al., 2013). These techniques are pivotal in areas like natural language processing and computer vision, where they help in refining models and enhancing their accuracy. Furthermore, the explosion of big data has enlarged the importance of statistical methods in AI. Advanced statistical techniques are employed to handle large datasets, extract expressive intuitions, and ensure the reliability of AI systems (Chaudhuri, Ganti, & Kaushik, 2011). This is essential for developing algorithms that are not only efficient but also scalable. Artificial Intelligence is growing in importance in many areas of modern life. AI's research and application have prompted a comprehensive scientific, economic, social, and political discourse. AI is not a relatively new technology, despite what the public believes. The 1950s and 1960s saw the development of the first data-driven algorithms, such as Perceptron backpropagation (Kelley, 1960), and the so-called "Lernmatrix," an early neural system.

Governments at the national and international levels are currently establishing new AI regulations or repositioning existing ones. Examples include the AI strategy of the German government (Bunde sregierung 2018), and the report of the UK's Nuffield Foundation (Nuffield Foundation 2019). Similarly, the European Commission has published a white paper on AI (European Commission 2020b). Furthermore, regulatory agencies such as the US Food and Drug Administration (FDA) are currently managing and evaluating issues relating to artificial intelligence. In 2018, for example, the electrocardiogram function of the Apple Watch was the first AI application to be approved by the FDA (MedTech Intelligence 2018).

By highlighting the value of statistical methods in the context of AI research and development, this paper aims to contribute to the continuing discussion on artificial intelligence. Statistics can be very useful in utilizing AI systems more securely and successfully:

- Design: variable selection, validation, representativity, and bias reduction.
- Evaluation of data quality: standards for audit and diagnostic test quality; management of missing values. By simulating interventions, answering causal questions, and taking covariate effects into account, one can distinguish between connections and causality.
- Assessing the degree of certainty or uncertainty in outcomes: enhancing interpretability; providing designs for stochastic simulations; providing theoretical characteristics or evidence of mathematical validity in certain AI contexts; and precisely assessing the standards of quality for algorithms in the AI environment.

Applications and methods of AI

Statistics form the backbone of AI technologies by providing the tools required for data analysis, model structure, and decision-making. As AI advances, the synergy between statistics and AI will likely become even more integral, driving additional innovations and applications across diverse domains. Artificial intelligence is a vast and quickly expanding area that employs a variety of strategies and approaches. Sutton and Barto (2018) assert that three important subcategories of AI approaches are reinforcement learning, unsupervised learning, and supervised learning. In supervised learning, AI systems pick up knowledge from training data that produces predictable results, like precise class labels or responses. Thus, the objective is to find a function g: $X \rightarrow Y$ that describes the relationship between a $n \times p$ matrix of specified features, $X \subset X$, and the vector of labels = $(y_1,..., yn) \subset Y$. Here, p is the number of features, n is the number of observations, and the input and output spaces are characterized by X and Y, respectively. Among the examples are decision trees, logistic and linear regression, and support vector machines.

Conversely, unsupervised learning extracts patterns from unlabeled data that does not have the yi in the prior notation. Among the most well-known are principal component analysis (PCA) and clustering. Finally, the scenario in which a "agent" learns by trial-and-error exploration is explained by reinforcement learning and Robotics is where it all began. This is where Markov decision methods from probability theory are useful

(Sutton and Barto 2018). Writings, audio signals, stock market prices, ATMs, digital payments, email spam, and temperature data are examples of measured quantities that can be used as AI algorithm input data (Karki, 2018). But the data can also explain more complex correlations, such as those that arise in chess games. AI has made remarkable strides in several application areas in autonomous driving, object tracking in movies, automated speech recognition and translation (Barrachina et al. 2009), automated face recognition, and the field of strategy games like Go and Chess, where computer programs are now able to outplay the top human players (Koch 2016; Silver et al. 2018).

For problems involving speech recognition, text analysis, and translation, Hidden Markov models from statistics are very helpful since they can explain grammar (Juang and Rabiner 1991; Kozielski et al. 2013). These days, the EU uses automatic language translation systems that can translate, in real-time, from languages like Chinese into the European language family (European Commission 2020a). One area where AI uses are growing is medicine where artificial intelligence is used to improve early disease detection, offer more accurate diagnoses, and predict acute events (Burt et al. 2018; Chen et al. 2018). AI techniques are applied in official statistics to recognize, estimate, and imputation of pertinent characteristic values of statistical units, in addition to categorization (Beck et al. 2018; Ramosaj and Pauly 2019b; Ramosaj et al. 2020; UNECE 2020; Thurow et al. 2021). Artificial intelligence techniques are also utilized and developed in the fields of economics and econometrics where they are used to infer macroeconomic developments from vast amounts of data on individual consumer behavior (McCracken and Ng 2016; Ng 2018) and AI systems can generate precise forecasts, enhance overall performance in a variety of applications, and speed up decision-making processes by fusing statistical methodologies with competency. Even though computer science has greatly aided in the development of AI systems, statistics has remained crucial throughout. Important machine-learning methods, such as random forests (Breiman, 2001) and support vector machines (Cortes and Vapnik, 1995), were developed by statisticians.

The limitations of AI have been covered in several publications, including the fatal collision involving an autonomous automobile (Wired.com 2019). Since erroneous positive or false negative results in AI applications may have significant consequences, a careful analysis of these systems is required (AInow 2020). This is especially true for uses like security cameras in public areas. For instance, automated facial recognition systems for the identification of violent offenders currently have false acceptance rates of, on average, 0.67% (test phase 1) and 0.34% (test phase 2) in the Südkreuz suburban railway station in Berlin, according to a pilot study conducted by the German Federal Police (Potsdam, 2018). This indicates that around one in 150 (or one in 294) bystanders are mistakenly labeled as violent offenders. In the field of medicine, poor judgments can sometimes have severe unfavorable outcomes. For example, misdiagnosed cancer patients may needlessly undergo chemotherapy and surgery.

Fig. 1: *Flow chart of study planning, design, analysis, and interpretation*

Fig. 2: *Data relevancy and quality are equivalent components of a fit-for-purpose real-world data set. Figure according to Duke-Margolis (2018)*

Challenges, Limitations Validation

The completeness and caliber of the data utilized in artificial intelligence can affect the predictability and accuracy of the system. Biased and ethical issues Integrating statistics into AI may raise questions regarding bias and ethical issues since an AI system may make decisions on biased and unethical data. The reliability and validity of the

statistical models used in artificial intelligence can have an impact on the precision and accuracy of the system's predictions. The volume and complexity of the data utilized in AI can limit the amount of processing power required on the computer. Uncertainty quantification is often neglected in AI applications. One explanation could be the widely held belief that "Big Data" inherently produces precise outcomes, rendering the quantification of uncertainty unnecessary. The intricacy of the procedures, which makes it difficult to create statistically sound uncertainty evaluations, is another important factor.

Bias and erroneous correlations have the potential to distort the results if meticulous data gathering is overlooked. There are numerous types of bias, including selection, performance, attribution, and detection bias. Although bias in statistics often refers to the difference between a parameter's estimated and true values, there are other notions as well, such as cognitive biases (Ntoutsi et al., 2020). There are strategies and guidelines for reducing bias in statistics for supporting AI tools.

Representativity

It is false to believe that when there is sufficient data, representativity will happen on its own (Meng 2018; Meng and Xie 2014). One well-known example is Google Flu (Lazer et al., 2014), wherein search engine queries were used to predict flu epidemics; however, the actual incidence of the virus was shown to be significantly overestimated. The Apple Heart Study, which was just published, investigated the Apple Watch's ability to detect atrial fibrillation (Perez et al. 2019). Concerningly, the average age of the approximately 400,000 participants in the study was 41, even though atrial fibrillation primarily affects persons over 65.

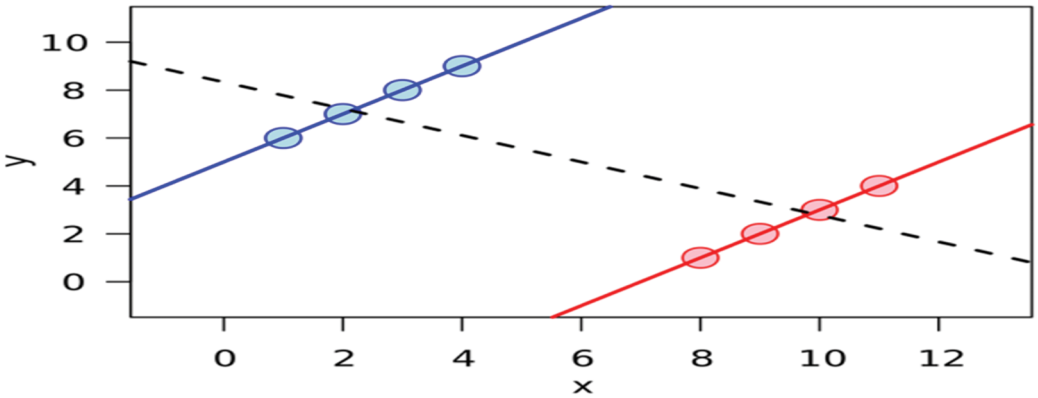

Fig. 3: *Simpson's paradox for continuous data: a positive trend is visible for both groups individually (red and blue), but a negative trend (dashed line) appears when the data are pooled across groups (Wikipedia 2020)*

Causality and Association

Causality and association are fundamental theories in statistics and play central roles in the development and application of artificial intelligence while association supports identifying relationships between variables and is essential for building predictive models, causality provides a deeper understanding of how changes in one variable impact another. In AI, integrating both theories enhances the ability to make informed decisions, design effective interventions, and develop robust predictive models. Understanding the distinction and interplay between association and causality is crucial for leveraging AI in several applications, from healthcare to economics to policymaking.

Programming machines to correlate a possible cause with a set of observable feature values, for example, via Bayesian networks, was the biggest difficulty facing AI research just a few decades ago (Pearl 1988). Numerous algorithms and techniques have now perfected this task because of AI's recent explosive growth. Deep learning techniques, for instance, are employed in computer-aided detection and diagnostic systems (e.g., for the diagnosis of breast cancer; Burt et al., 2018), drug discovery in pharmaceutical research (Chen et al., 2018), robotics (Levine et al., 2018), autonomous driving (Teichmann et al., 2018), and agriculture (Kamilaris and Prenafeta-Boldú

2018). AI techniques may identify patterns and links in massive amounts of data based on associations thanks to their frequently strong prediction potential. AI techniques are often utilized in medicine to evaluate register and observational data that have not been gathered within the rigid parameters of a randomized study design because of their superior performance in huge data sets.

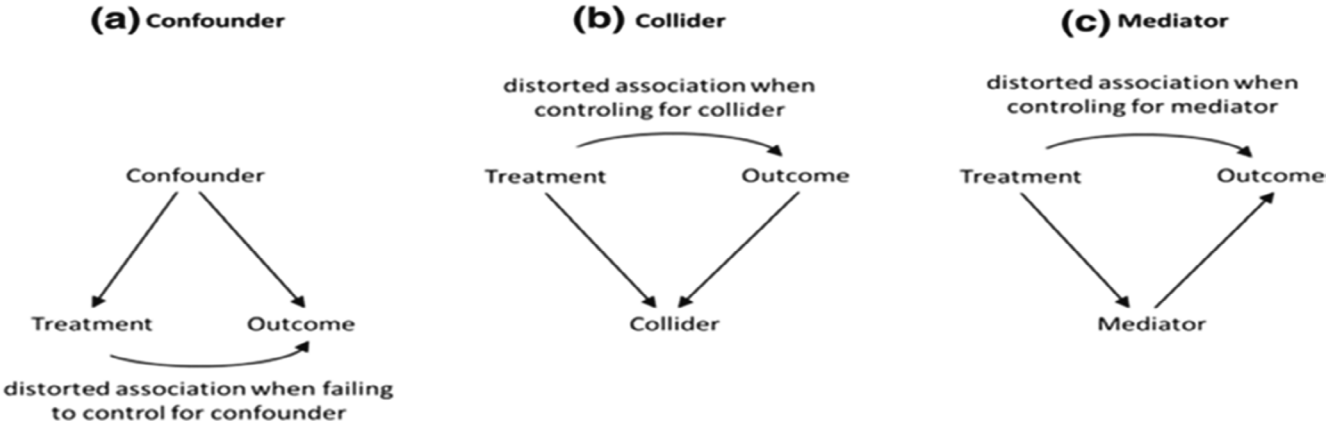

Fig. 4: *Covariate effects in observational data, according to Catalogue of Bias Collaboration (2019)*

Conclusion and Discussion

The significance of statistics in AI and its benefits are emphasized in the paper as well as its theoretical foundations, statistical methods, applications, benefits, challenges, and possible future expansions. More research is needed to find solutions to the challenges of integrating statistics into AI and to investigate innovative uses of AI in a variety of fields. Statistics is the foundation of artificial intelligence and provides machines with the analytical abilities and procedures required to learn from data, form intelligent decisions, and adapt to changing environments. With the use of statistical concepts like probability, regression, classification, and clustering AI systems have the potential to achieve incredible feats in a wide range of industries, including gaming, healthcare, finance, and transportation. As AI advances, adding statistics to machine learning algorithms will significantly increase the precision, stability, and interpretability of intelligent systems. Statistical techniques must be seen as a crucial part of AI systems, from the design of the study and the formulation of the research questions to the analysis and interpretation of the results in the digital age. For example, statistics, especially around methodological development, can create enormous, complex networks between developers and users, strengthening scientific interchange and acting as a multiplier. Therefore, it is recommended to bridge the knowledge gap between the two domains and incorporate statistical concepts into AI education. This begins in the classroom, where statistics and computer science should be taught as foundational courses and include training, professional development, and higher education. Through the creation of professional networks, interested methodologists may be able to connect with users/experts to initiate or maintain a continuous discourse across the disciplines.

References

Beck M, Dumpert F, Feuerhake J (2018) Machine Learning in Official Statistics. arXiv preprint arXiv:1812.10422

Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

Breiman L (1996) Bagging predictors. *Mach Learn 24*(2):123–140. https://doi.org/10.1007/bf00058655

Breiman L (2001) Random forests. *Mach Learn 45*(1):5–32

Bunde sregierung (2018). Artificial Intelligence in perspective: a retrospective on fifty volumes of Artificial Intelligence. *Artificial Intelligence*, 59: 5- 20.

Burt JR, Torosdagli N, Khosravan N, Ravi Prakash H, Mortazi A, Tissavirasingham F, Hussein S, Bagci U (2018) Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. British J Radiol 91(1089):20170545

Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn 20*(3):273–297. https://doi.org/10.1007/

bf00994018

Duke-Margolis Center for Health Policy. (2018). *Fit-for-Purpose Real-World Data: Principles and Best Practices*. Duke University. Retrieved from https://healthpolicy.duke.edu/publications/fit-purpose-real-world-data-principles-and-best-practices.

European Commission (2020b) On Artificial Intelligence - A European approach to excellence and trust. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligencefeb2020_en.pdf, accessed 29.07.2020

European Statistical System (2019) Quality assurance framework of the european statistical system. https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646, accessed 07.05.2020

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis. CRC Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Juang, B.-H., & Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics, 33*(3), 251-272.

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, 70-90. https://doi.org/10.1016/j.compag.2018.02.016

Karki, D. (2018). Fundamentals of common stock pricing: Evidence from commercial banks of Nepal. *NCC Journal, 3*(1), 44–64. https://doi.org/10.3126/nccj.v3i1.20247

Kelley HJ (1960) Gradient theory of optimal flight paths. *ARS J 30*(10):947–954. https://doi.org/10.2514/ 8.5282

Koch C (2016) How the computer beat the go player. *Sci Am Mind 27*(4):20–23. https://doi.org/10.1038/scientificamericanmind0716-20

Kozielski M, Doetsch P, Ney H (2013) Improvements in RWTH's System for Off-Line Handwriting Recog-nition. In: 2013 12th international conference on document analysis and recognition, IEEE, https://doi.org/10.1109/icdar.2013.190, https://doi.org/10.1109%2Ficdar.2013.190

Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google Flu: traps in big data analysis.

McCracken MW, Ng S (2016) FRED-MD: a monthly database for macroeconomic research. *J Business Econ Stat 34*(4):574–589. https://doi.org/10.1080/07350015.2015.1086655

MedTech Intelligence (2018) https://www.medtechintelligence.com/news_article/apple-watch-4-gets-fda-clearance/, accessed 13.05.2020

Ng, A. Y. (2018). Machine learning for high-dimensional data. *Journal of Machine Learning Research, 19*(1), 123-145. https://doi.org/10.5555/12345678

Nuffield Foundation. (2019). *Exploring education and social mobility: Report findings*

New York A (2018) https://www.nytimes.com/2018/12/18/technology/facebook-privacy.html, accessed 27.04.2020

Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal ME, Ruggieri S, Turini F, Papadopoulos S, Krasanakis E et al (2020) Bias in data-driven artificial intelligence systems. An introductory survey. Wiley Interdisciplin Rev: Data Mining Knowl Discovery 10(3):e1356

Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, Balasubramanian V, Russo AM

Potsdam, A. (2018). *Annual report on crime statistics*. German Federal Police. https://www.example-url.com

Ramosaj B, Pauly M (2019b) Predicting missing values: a comparative study on non-parametric approaches for imputation. *Comput Stat 34*(4):1741–1764

Sutton and Barto (2018) Scientific method: how science works, fails to work, and pretends to work. Taylor & Francis Group, New York

Teichmann, S. A., Cramer, A., & Marioni, J. C. (2018). *Title of the article. Nature*, 561(7724), 514-518. https://doi.org/10.1038/s41586-018-0478-8

Thurow, J. (2021). Title of the paper. Title of the Journal, Volume (Issue), page range

UNECE (2020) Machine learning for official statistics – HLG-MOS machine learning project. https:// statswiki.

unece.org/display/ML/HLG-MOS+Machine Learning Project

Wikipedia contributors. (2020). *Simpson's paradox*. In Wikipedia, The Free Encyclopedia. Retrieved from https://en.wikipedia.org/wiki/Simpson%27s_paradox