S.Ghosh, Vol 1(2), pp 86-97
https://doi.org/10.55306/CJDTES.2024.1204
e-ISSN: 3048-8389
Ci-STEM Journal of Digital Technologies and Expert Systems

Study of Machine Learning Algorithms: A Historical Perspective

Sushmita Ghosh
Board of Practical Training (Eastern Region), (An Autonomous Organization under Department of Higher Education, Ministry of Education, Govt. o India) Kolkata, West Bengal, India
email: sghosh@bopter.gov.in

Open Access
Review Article
Received : 17/07/2024
Accepted : 28/10/2024
Published : 06/11/2024
Corresponding author email:
sghosh@bopter.gov.in
Citation:
S. Ghosh, "Study of Machine Learning Algorithms: A Historical Perspective," Ci-STEM Journal of Digital Technologies and Expert Systems, Vol. 1(2), pp. 86-97, 2024, doi: 10.55306/CJDTES.2024.1204

Published By:
Ci-STEM Global Services Foundation, India.

Abstract:
ML stands for machine learning, a groundbreaking technological paradigm that has transformed the way we approach diverse areas of society from clinical diagnostics to algorithmic trading, from personalized media delivery to self-driving cars. In this article, we will look into the history of ML algorithms — in the order of their chronology. This can aid in providing an extensive perspective of how various ML techniques evolved, the theoretical underpinnings that strengthen them and the potential of forthcoming research in this branch of computational intelligence. Tracing this history helps us to appreciate how the interplay between theoretical innovations, the expanding availability of computational resources, and the demands of complex real-world problems have shaped the field of ML. From a few decades ago to today, this evolution tells the story of how ML has become such an influential driver of much of the technology that we see today. Lastly, the article discusses the interaction among improvements in algorithm design, computational infrastructure, and challenges brought by data tasks, paving the way for more advances in the years to come.
Keywords: *Computational intelligence, Historical development, Machine Learning (ML) Algorithms, Real-world applications, Research directions, Theoretical foundations.*

1. INTRODUCTION

The most prominent sub-discipline of artificial intelligence (AI) is machine learning (ML); this field is primarily focused on the formulation of algorithms that enable machines to learn from data and predict new outcomes without explicit engineering procedures. Several iterations of machine learning have occurred over years, wherein simple statistical techniques based on linear models have morphed into advanced machine learning algorithms, quickly overtaking the classics and evolving into deep learning networks that leverage multi-layered neural networks and perform stupendously. Based on work recorded in [1], this article outlines a brief historical survey of machine learning algorithms, with the aim of mapping out a timeline of key innovations and technological developments to demonstrate the vast range of algorithmic breadth and methodological depth associated with the evolution of this fast-moving field. As this historical overview demonstrates, we have moved from knowledge-engineered systems to data-driven inductive learning paradigms that form the foundation of a large number of intelligent applications today.

2. THE PRE 1950S — EARLY FOUNDATIONS OF MACHINE LEARNING

Long before the discipline of machine learning was formally established, early research laid the foundations of contemporary ML algorithms:

Opus on ML: Machine Learning: Tubes of Pre-Learning Data (Pre 1950)

While machine learning was only officially outlined as a standalone field of study in the mid-late 20th Century, it built heavily on earlier research across numerous fields that established key concepts and methodologies in pursuit of increasing the efficacy of modern-day machine learning algorithms.

These early lines of thought, emerging chiefly from statistical techniques and the then-nascent field of cybernetics, contributed vital building blocks in the future development of machine intelligence.

Statistical Methods in the 19th Century: The Genesis of Regression and Parameter Estimation

Historical Context and Development

The foundations of modern machine learning originated in 19th-century mathematical statistics, particularly through the development of regression analysis and parameter estimation techniques [2]. The period marked a crucial transition from descriptive to inferential statistics, establishing mathematical frameworks that would later become fundamental to machine learning.

The Method of Least Squares

Historical Development

The Method of Least Squares emerged through parallel developments by two mathematical giants:

1. Carl Friedrich Gauss (1795-1799): Developed the method while studying astronomical data [3]
2. Adrien-Marie Legendre (1805): First published formal description in "Nouvelles méthodes pour la détermination des orbites des comètes" [4]

Mathematical Formulation

Core Concept

The method addresses the fundamental problem of fitting mathematical models to empirical data through optimization.

Mathematical Framework

1. **Given**:
   - Data points: $(x_i, y_i)\ for\ i = 1, \dots n$
   - Model function: $f(x, \beta)$
   - Parameters: $\beta = (\beta_1, \dots \dots, \beta_k)$
2. **Objective Function**: $S(\beta) = \sum_{i=1}^{n} [y_i - f(x_i, \beta)]^2$
3. **Optimization Condition**: $\frac{\partial S}{\partial \beta_j} = 0\ for\ j - 1, \dots ., k$

Theoretical Implications

The method established several fundamental principles [5]:

1. **Error Minimization**:
   - Systematic approach to handling observational errors
   - Foundation for modern optimization techniques
2. **Parameter Estimation**:
   - Maximum likelihood estimation under normal errors
   - Basis for modern statistical inference
3. **Model Fitting**:
   - Framework for relationship quantification
   - Precursor to modern regression analysis

Impact on Modern Machine Learning

The 19th-century developments in statistical methods directly influenced modern machine learning through:

1. **Optimization Framework**
   - Gradient-based optimization methods
   - Error minimization principles
   - Loss function design
2. **Model Evaluation**
   - Residual analysis
   - Goodness-of-fit measures
   - Validation techniques

Cybernetics (1940s-1950s): Understanding Machine Intelligence & Self-Regulation

Cybernetics, a kind of transdisciplinary discipline, came into being in the mid-20th century and had a profound impact on the early thinking of AI and ML. Cybernetics, initially, was pioneered by people like Norbert Wiener, and greatly influenced by John von Neumann's work, they explored the pessimistic principles of control and communication in biological organisms and engineered systems [6] . Norbert Wiener's Cybernetics: or Control and Communication in the Animal and the Machine (1948) introduced the foundational concepts, focusing on feedback loops, information theory, and homeostasis that enables goal-directed behavior and self-regulation in complex systems [7]. Wiener proposed that intelligence emerges from well-constructed feedback loops and information processing which do not depend on the biological or artificial nature of the underlying components. While John von Neumann made significant contributions to fields like computer architecture and game theory, he also had interactions with cybernetics, notably in developing self-reproducing automata and logical constructs for designing complex self-organizing systems [8]. Such investigations into the principles of feedback control, information processing and self-organization within the framework of cybernetics resulting in important conceptual groundings for machine learning. Rather than purely symbolic or rule-based approaches, which would come to define early symbolic AI, they looked at intelligence as a process of computation that responds to feedback and makes use of mechanisms for adaptation and learning. Indeed, the cybernetic view highlighted that systems interact dynamically with their environment, shaping the quality of that interaction by learning from experience via feedback, principles well embedded in contemporary paradigms in machine learning, including reinforcement learning and adaptive control systems.

3. THE EMERGENCE OF MACHINE LEARNING (1950S–1970S)

"Machine learning" was originally a term created in the 1950s just as the formulation of the field as a discipline began. During this time, several key algorithms and paradigms were developed that would set the stage for future developments in the field.

1958: Perceptron – The Birth of Artificial Neural Nets

The Perceptron was invented by Frank Rosenblatt in 1958 and is considered as one of the earliest instantiations of an ANN [9]. The Perceptron marked a significant step in simulating information processing mechanisms similar to the human brain. The Perceptron is a linear classifier that can solve binary classification problems. The perceptron works by calculating the weighted sum of the input features, adding a bias term to it, and then passing the sum through an activation function — for example, a Heaviside step function [10]. The Perceptron learning rule, or the learning algorithm by which the Perceptron learns, is an online learning algorithm that updates the weights and bias iteratively whenever a training example is misclassified. This rule adjusts the weight according to both: the input vector to the neuron and the error signal, so that it converges towards a decision boundary that accommodates the classes: [11] So, although it could only represent linearly separable problems [10] the Perceptron was a very significant proof-of-concept and an inspiration for more advanced neural network architectures that followed.

Bayesian Inference (1960s): Reasoning with probabilities and combining evidence

The Bayesian approach to learning arose in the 1960s as a powerful paradigm for probabilistic reasoning and inference in machine learning. A Bayesian approach assumes Bayes' Theorem, an essential rule of probability theory, which is then used to update one's beliefs (prior probabilities) based on new evidence to determine posterior probabilities [12]. Bayesian inference allows us to compute the posterior probability of a hypothesis (class label, model parameter, etc) based on observed data in machine learning contexts. Based on this framework, probabilistic classifiers were created, such as Naive Bayes classifiers, which assume conditional independence of features given the class label [13]. Even though it oversimplifies, Naive Bayes classifier was very successful in a number of tasks, especially with text classification and spam filtering, thanks to its simplicity, high computational speed and high-dimensional data positive performance [14]. Unlike purely deterministic approaches, the Bayesian paradigm offered a principled framework to address uncertainty and incorporate prior knowledge into the learning process.

Early Decision Trees (1960s): Recursive Partitioning for Data Classification

The study of the decision trees was an independent line of research emerging around this time, focusing on hierarchical partitioning of data for classification and prediction. ID3 (short for Iterative Dichotomiser 3) was one of the early influential decision tree algorithms, proposed by Ross Quinlan [15]. ID3 is a top-down, greedy algorithm that constructs a decision tree by recursively splitting the data set based on feature values. Each node uses the information gain metric to select the splitting feature; it gauges how much entropy (impurity) of the target variable is reduced after parting the data according to the specified feature [16]. In this context, entropy measures the uncertainty of the classification label distribution. While building a tree, ID3 uses information gain in selecting a feature that maximizes information gain, it creates internal nodes of the tree, while branches hold the feature values, and leaf nodes represent class labels. Although early decision tree algorithms, such as ID3, had limitations in their ability to handle continuous features and were prone to overfitting [17], they provided the conceptual basis for advanced approaches to decision trees such as C4. 5 and CART, which were progressively accepted for their interpretability and non-parametric nature.

4. THE ORIGINS OF SYMBOLIC AND CONNECTIONIST PARADIGMS (1970S-1980S)

The 1970s and 1980s were a key period in the development of Artificial Intelligence (AI) with dramatic improvements across both symbolic and connectionist modeling methods. The neural networks' interest revived in dominion due to an algorithmic revolution that happened in this epoch, as rule-based systems began to grow into employment.

Use the following to prove your familiarity with the topic: Expert Systems (1970s): Rule-Based Reasoning and Knowledge Engineering

Expert systems emerged as a dominant paradigm within symbolic AI in the 1970s. These systems were based on the principle of rule-based reasoning, using an explicit representation of domain-specific knowledge to replicate the problem-solving behaviors of human experts [18]. Expert systems initially relied on knowledge representation formalisms, such as production rules (IF-THEN statements), semantic networks, and frames, to structure and encode expert knowledge knowledge in a declarative way [19]. Early algorithmic development in this area focused predominantly on the construction of efficient mechanisms for inference, such as forward chaining and backward chaining algorithms, to derive conclusions and recommendations from the encoded knowledge base [20]. These systems employed formal logic and heuristic search methods to mimic expert as well as human reasoning across a wide range of knowledge domains; MYCIN [21] for medical diagnoses and DENDRAL [22] for chemical analysis, for example. Towards the end of the decade, it became clear that expert systems have serious limitations, such as the knowledge acquisition bottleneck and brittleness, that is, inability to respond to unseen situations not included in the original rule set [23].

Although many papers were published with a connectionist flavour during these years, a major breakthrough that led to the renewal of connectionist models in 1986 was the formalization of backpropagation. Backpropagation, introduced by Rumelhart, Hinton and Williams, offered a computation tractable way to efficiently find the weights for multi-layered artificial neural networks [24]. This algorithm was an elegant solution to the challenge of credit assignment problem for networks with hidden layers, allowing the optimizations of the network weights to minimize a predefined error function. Backpropagation combines the principles of gradient descent with the chain rule of calculus to progressively calculate the gradient of the error function regarding each weight in the network. The gradients for each layer are propagated backwards through the network so that the weights can iteratively be adjusted to minimize the error and improve the performance of the network on the task in question [25]. However, the real breakthrough came with the introduction of the back-propagation algorithm, which addressed the earlier inability of multilayer networks to be trained because of the absence of efficient training algorithms, igniting a new wave of research activity in connectionist AI and neural networks.

The new and powerful Support Vector Machine (SVM) algorithm originated in the 1980s, when Vladimir Vapnik and Alexey Chervonenkis introduced the concept based on statistical learning theory [26]. The Support Vector Machines or SVM main purpose is either classification and regression task, mainly used in high-dimensional space. SVM's fundamental idea is to find an optimal hyperplane that

separates the data of different classes in such a way as to create the widest gap possible between the data points of different classes in the feature space. The ideal hyperplane is determined by maximizing the margin, which is the distance between hyperplane and the closest data points from each class (this is called the support vectors) [27]. SVMs are based on structural risk minimization, which aims to minimize generalization error instead of minimizing empirical error on training data only [28]. In addition, the concept of kernel trick paved the way for SVMs to tackle a variety of non-linear classification challenges by implicitly projecting input data into high-dimensional feature spaces using kernel functions without needing to directly compute the transformations [29]. SVMs were thus established as the algorithm of choice for many machine learning tasks in the decades to follow due to these characteristics along with the many advantageous theoretical properties that accompanied the formulation.

5. MACHINE LEARNING GOES MAINSTREAM (1990S):

The 1990s represented a critical period in machine learning history, marked by significant algorithmic breakthroughs and practical applications. This era saw the transition of machine learning from purely academic research to industrial applications, driven by increased computational capabilities and algorithmic innovations. Key Algorithmic Developments are given below.

Random Forests (1995)

Breiman [30] introduced Random Forests as an ensemble learning method, building upon earlier work in bagging predictors [31]. The algorithm's key technical components include:

1. **Bootstrap Aggregating (Bagging)**
   - Random sampling with replacement from training data
   - Typically 63.2% of original samples selected for each tree
   - Remaining 36.8% used as Out-of-Bag (OOB) samples for error estimation
2. **Random Feature Selection**
   - At each split, randomly select $m \leq p$ features (where p is total features)
   - For classification: $m = \sqrt{p}$ (default)
   - For regression: $m = p/3$ (default)
   - Decorrelates trees, reducing variance (Liaw & Wiener, 2002)
3. **Tree Construction Protocol**
   - CART (Classification and Regression Trees) methodology
   - Gini impurity or entropy for classification
   - Mean Squared Error (MSE) for regression
   - No pruning of fully grown trees

K-Nearest Neighbours (K-NN)

While originally proposed by Fix and Hodges [32], K-NN saw significant theoretical development and practical implementation in the 1990s [33]:

1. **Distance Metrics**
   - Euclidean distance (most common)
   - Manhattan distance
   - Minkowski distance
   - Mahalanobis distance for correlated features
2. **Algorithm Optimizations**
   - Ball trees and KD-trees for efficient nearest neighbor search
   - Approximate nearest neighbor methods
   - Complexity reduction from O(nd) to O(log n) for queries
   - Local weight functions[34].

Clustering Algorithms

K-means Clustering

MacQueen [35] developed the original algorithm, but the 1990s saw significant improvements [36]:

1. **Technical Components**
   - Objective function: minimize within-cluster sum of squares
   - Lloyd's algorithm for optimization
   - K-means++ initialization [37]
   - Complexity: O(tknd), where:
     - t = iterations
     - k = clusters
     - n = samples
     - d = dimensions
2. **Variants Developed**
   - Fuzzy K-means
   - Kernel K-means
   - Mini-batch K-means

Hierarchical Clustering

Major developments in hierarchical clustering methods [38]:

1. **Linkage Criteria**
   - Single linkage: min distance
   - Complete linkage: max distance
   - Average linkage: mean distance
   - Ward's method: minimum variance

Hidden Markov Models (HMMs)

Rabiner [39] provided the foundational framework, with significant developments in the 1990s:

1. **Technical Components**
   - State transition probability matrix A
   - Observation probability matrix B
   - Initial state distribution π
   - Forward-backward algorithm for parameter estimation
   - Viterbi algorithm for optimal state sequence
2. **Training Algorithms**
   - Baum-Welch algorithm (EM for HMMs)
   - Segmental K-means
   - Maximum Mutual Information Estimation (MMIE)

6. THE DEEP LEARNING REVOLUTION (2000S-PRESENT)

The 2000s saw a significant shift toward deep learning techniques, driven by increased computational power, large datasets, and advances in neural network architectures:

Early Foundations (2000-2009) Key Algorithms and Architectures

1. **Convolutional Neural Networks (CNNs)**
   - LeNet-5 architecture pioneered the fundamental CNN principles as demonstrated by LeCun et al. [40] in their seminal work on document recognition
   - Key components: Convolution layers, pooling layers, fully connected layers
   - Primarily used for handwritten digit recognition, achieving state-of-the-art performance on the MNIST dataset
   - Innovation: Local receptive fields and weight sharing, which LeCun et al. [41] later identified as crucial for deep learning success

2. **Deep Belief Networks (DBNs)**
    - Revolutionized deep learning through Hinton et al.'s [42] breakthrough work on fast learning algorithms
    - Layer-wise pretraining using Restricted Boltzmann Machines (RBMs)
    - Unsupervised learning followed by supervised fine-tuning
    - Breakthrough: Solved vanishing gradient problem in deep networks, as detailed in the original paper
3. **Restricted Boltzmann Machines (RBMs)**
    - Binary stochastic units with symmetric connections, fundamental to Hinton's work [42].
    - Training using Contrastive Divergence algorithm
    - Used for dimensionality reduction and feature learning
    - Foundation for deep belief networks

The Renaissance Period (2010-2014): Revolutionary Algorithms

1. **AlexNet Architecture**
    - Krizhevsky et al. [43] introduced this groundbreaking architecture that won the ImageNet competition
    - 8 layers (5 convolutional, 3 fully connected)
    - ReLU activation functions, which proved crucial for training deep networks
    - Local Response Normalization
    - Dropout regularization, significantly reducing overfitting
2. **Word2Vec**
    - Mikolov et al. [44] developed this influential word embedding technique
    - Continuous Bag of Words (CBOW) and Skip-gram models
    - Negative sampling optimization
    - Created dense vector representations of words
3. **Generative Adversarial Networks**
    - Goodfellow et al. [45] introduced the revolutionary GAN framework
    - Generator and Discriminator networks in an adversarial setting
    - Applications: Image generation, style transfer
    - Innovation: Minimax optimization framework

7. LARGE LANGUAGE MODELS

Large Language Models (LLMs) have revolutionized natural language processing and artificial intelligence. These models, based on transformer architectures [45], have demonstrated unprecedented capabilities in language understanding, generation, and complex reasoning tasks. This technical document analyzes current LLMs, their architectures, capabilities, and limitations.

Fundamental Architecture

Transformer-Based Models

The foundation of modern LLMs stems from the transformer architecture introduced by[45]. Key components include:

- Multi-head self-attention mechanisms
- Position-wise feed-forward networks
- Layer normalization
- Residual connections

This architecture enabled parallel processing and better handling of long-range dependencies compared to previous RNN-based approaches [45].

Major Model Families

GPT Series

The GPT (Generative Pre-trained Transformer) family, developed by OpenAI, has been instrumental in scaling language models:

1. **GPT-3 (175B parameters)**
    - Introduced few-shot learning capabilities [46]
    - Demonstrated emergent abilities with scale
    - Used dense transformer decoder architecture
    - Training approach: Causal language modeling
2. **GPT-4**
    - Multimodal capabilities (OpenAI, 2023)
    - Enhanced reasoning and domain expertise
    - Improved factual accuracy and reduced hallucinations
    - Architecture details partially undisclosed

PaLM Models

Google's Pathways Language Model series:

1. **PaLM (540B parameters)**
    - Scaled transformer architecture [46].
    - Demonstrated chain-of-thought reasoning
    - Advanced multilingual capabilities
    - Pathways system for efficient training
2. **PaLM 2**
    - Enhanced multilingual understanding
    - Improved mathematical and reasoning capabilities
    - More efficient architecture than predecessor
    - Focus on responsible AI development

BERT and Variants

1. **BERT**
    - Bidirectional encoder representations [45]
    - Masked language modeling objective
    - Next sentence prediction task
    - Transformer encoder architecture
2. **RoBERTa**
    - Modified BERT training approach [47]
    - Removed next sentence prediction
    - Dynamic masking
    - Larger batch sizes and training data

Technical Innovations: Scaling Techniques

1. **Mixture of Experts**
    - Sparse activation of model parameters
    - Improved computational efficiency
    - Router-based expert selection
    - Demonstrated in Switch Transformers [48]
2. **Parameter-Efficient Fine-tuning**
    - LoRA: Low-rank adaptation [49]
    - Prompt tuning and soft prompts
    - Adapter layers
    - Reduced memory and computational requirements

Training Optimizations

1. **Pre-training Strategies**
    - Masked language modeling
    - Causal language modeling
    - Contrastive learning approaches
    - Multi-task pre-training objectives
2. **Architectural Improvements**
    - Flash attention for efficiency [50]
    - Rotary positional embeddings
    - Grouped-query attention
    - Memory-efficient optimizations

Current Challenges and Solutions

Training Efficiency

- Gradient accumulation techniques
- Mixed precision training
- Pipeline parallelism
- Tensor parallelism [51]

Model Limitations

1. **Hallucination Mitigation**
    - Retrieval-augmented generation
    - Constitutional AI approaches
    - Fact-checking mechanisms
    - Uncertainty quantification
2. **Computational Resources**
    - Quantization techniques
    - Pruning methods
    - Distillation approaches
    - Efficient inference strategies

8. NEW DIRECTIONS AND CURRENT TRENDS

With the continuous advancement of machine learning, several emerging trends and focal points have been defining in shaping the present scenario:

- Federated Learning: Train models in a decentralized manner across devices while keeping the data on the device.
- AutoML (Automated machine learning): AutoML is designed to automate the selection and hyper parameter optimization for machine time an.
- Explainable AI (XAI): As ML models grow increasingly complex, transparency and interpretability will be in high demand. XAI was developed to create methods that explain how models make decisions, particularly in sensitive applications such as healthcare and finance.
- Ethics and Bias in AI: While ML systems are deployed to real world applications, concerns about fairness, transparency and bias have come up. One of the major open research areas is to ensure that the algorithms are ethical and not biased.

9. CONCLUSION & FUTURE WORK

Machine learning algorithms have experienced rapid growth and radical paradigm shifts between the time when they were conceived and the present. With the evolution of statistical methods such as linear models, neural networks, reinforcement learning and computational techniques, the field has witnessed a progressive trajectory, driven towards tackling more complex and difficult computational problems with accompanying computational infrastructure advancements. The development in machine learning has often been punctuated by individual milestone algorithm discoveries and impactful model instantiations, which together served to propel the field into new research directions.

Examples of such advancements are the kernel methods, probabilistic graphical models, and distributed learning algorithms, each of which has enabled the creation of more capable ML systems. Pondering the roads ahead, the potential fusion of Artificial Intelligence with avant-garde technologies—from quantum computing architectures to neuromorphic processing environments—is anticipated to significantly transform the fabric of machine learning models, potentially pushing the limits of what we consider computationally feasible, and unlocking novel vistas of previously unapproachable problem sets. The continuous development of ML techniques points to an increasingly common paradigm where ML algorithms experience greater adaptability, efficiency, and cognition-mimicking abilities.

New Directions and Current Trends

With the continuous advancement of machine learning, several emerging trends and focal points have been defining in shaping the present scenario:

- Federated Learning: Train models in a decentralized manner across devices while keeping the data on the device.
- AutoML (Automated machine learning): AutoML is designed to automate the selection and hyper parameter optimization for machine time an.
- Explainable AI (XAI): As ML models grow increasingly complex, transparency and interpretability will be in high demand. XAI was developed to create methods that explain how models make decisions, particularly in sensitive applications such as healthcare and finance.
- Ethics and Bias in AI: While ML systems are deployed to real world applications, concerns about fairness, transparency and bias have come up. One of the major open research areas is to ensure that the algorithms are ethical and not biased.

DECLARATIONS:

Acknowledgments	**:**	Not applicable.
Conflict of Interest	**:**	The author declares that there is no actual or potential conflict of interest about this article.
Consent to Publish	**:**	The authors agree to publish the paper in the Global Research Journal of Social Sciences and Management.
Ethical Approval	**:**	Not applicable.
Funding	**:**	Author claims no funding was received.
Author Contribution	**:**	Both the authors confirms their responsibility for the study, conception, design, data collection, and manuscript preparation.
Data Availability Statement	**:**	The data presented in this study are available upon request from the corresponding author.

REFERENCES

[1] X. Zhang, F. Guo, T. Chen, L. Pan, G. Beliakov, and J. Wu, "A Brief Survey of Machine Learning and Deep Learning Techniques for E-Commerce Research," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 18, no. 4, pp. 2188–2216, Dec. 2023, doi: 10.3390/jtaer18040110.

[2] S. M. Stigler, "Gauss and the Invention of Least Squares," *The Annals of Statistics*, vol. 9, no. 3, May 1981, doi: 10.1214/aos/1176345451.

[3] C. F. Gauss, *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Cambridge University Press, 2011. doi: 10.1017/CBO9780511841705.

[4] Adrien Marie LEGENDRE, *Nouvelles methodes pour la determination des orbites des cometes*, 1st ed. Firmin Didot, Paris, 1806.

[5] Stigler Stephen M, *The History of Statistics: The Measurement of Uncertainty before 1900*, 1st ed. President and Fellows of Harvard College, 1986.

[6] Ronald R. Kline, *The Cybernetics Moment: Or Why We Call Our Age the Information Age. (New Studies in American Intellectual and Cultural History.)*. Johns Hopkins University Press, Baltimore, 2015.

[7] Norbert Wiener, *Cybernetics or Control and Communication in the Animal and the Machine*. The MIT Press, 2019.
[8] John Von Neumann, *Theory of Self-Reproducing Automata*, vol. 1. University of Illinois Press, 1966.
[9] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.," *Psychol Rev*, vol. 65, no. 6, pp. 386–408, 1958, doi: 10.1037/h0042519.
[10] Marvin Minsky and Seymour A. Papert, *Perceptrons – An Introduction to Computational Geometry Exp Ed: An Introduction to Computational Geometry*. MIT Press, 1988.
[11] Simon Haykin, *Neural Networks: A Comprehensive Foundation*. Macmillan USA , 1994.
[12] F. R. S. Bayes and Price, "An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S," *Philos Trans R Soc Lond*, vol. 53, pp. 370–418, Dec. 1763, doi: 10.1098/rstl.1763.0053.
[13] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," *Mach Learn*, vol. 29, no. 2/3, pp. 103–130, Nov. 1997, doi: 10.1023/A:1007413511361.
[14] Jason D. Rennie, Lawrence Shih, Jaime Teevan, and David Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," in *Proceedings of the Twentieth International Conference on Machine Learning*, Jul. 2003.
[15] J. R. Quinlan, "Induction of decision trees," *Mach Learn*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.
[16] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.
[17] Tom Mitchell, *Machine Learning*. McGraw Hill, 1997.
[18] Edward A. Felgenbaum, "The Art of Artificial Intelligence: Themes and Case Studies of Knowledge Engineering," in *IJCAI'77: Proceedings of the 5th international joint conference on Artificial intelligence*, Aug. 1977, pp. 1014–1029.
[19] Ronald J. Brachman and Hector J. Levesque, *Readings in Knowledge Representation*. Morgan Kaufmann Publishers, 2000.
[20] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*. Pearson, 2009.
[21] Bruce G. Buchanan and Edward H. Shortliffe, *Rule-based Expert Systems : The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, 1984.
[22] Robert K. Lindsay, Bruce G. Buchanan, Edward A. Feigenbaum, and Joshua Lederberg, *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*. McGraw-Hill Companies, 1980.
[23] Frederick Hayes-Roth, Donald A. Waterman, and Douglas B. Lenat, *Building expert systems*. Addison-Wesley Longman Publishing Co., Inc., 1983.
[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.
[25] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[26] Vladimir N. Vapnik, *Statistical Learning Theory Vladimir N. Vapnik* . Wiley-Interscience, 1998.
[27] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
[28] M. Pastore, P. Rotondo, V. Erba, and M. Gherardi, "Statistical learning theory of structured data," *Phys Rev E*, vol. 102, no. 3, p. 032119, Sep. 2020, doi: 10.1103/PhysRevE.102.032119.
[29] Aizerman. M.A, Braverman. E.M, and Rozonoer. L.I., "Theoretical Foundations of Potential Function Method in Pattern Recognition," *Automation and Remote Control*, vol. 25, pp. 917–936, 1964.
[30] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[31] L. Breiman, "Bagging predictors," *Mach Learn*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/BF00058655.

[32] E. Fix and J. L. Hodges, "Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties," 1951.
[33] Belur. V. Dasarathy, "Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques," *IEEE Computer Society Press, Los Alamitos,* 1991.
[34] S. A. Dudani, "The Distance-Weighted k-Nearest-Neighbor Rule," *IEEE Trans Syst Man Cybern*, vol. SMC-6, no. 4, pp. 325–327, Apr. 1976, doi: 10.1109/TSMC.1976.5408784.
[35] J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probabilit*, 1967, pp. 281–297.
[36] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990. doi: 10.1002/9780470316801.
[37] David Arthur and Sergei Vassilvitskii, "K-Means++: The Advantages of Careful Seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA*, Jan. 2007, pp. 7–9.
[38] S. C. Johnson, "Hierarchical Clustering Schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, Sep. 1967, doi: 10.1007/BF02289588.
[39] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989, doi: 10.1109/5.18626.
[40] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.
[41] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[42] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Comput*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006, doi: 10.1162/neco.2006.18.7.1527.
[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[44] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS'13: Proceedings of the 27th International Conference on Neural Information Processing System*, 2013, pp. 3111–3119.
[45] Ian J. Goodfellow *et al.*, "Generative Adversarial Networks," in *Proceedings of the International Conference on Neural Information Processing Systems* , 2014, pp. 2672–2680.
[46] Tom B. Brown, Benjamin Mann, and Nick Ryder, "Language Models are Few-Shot Learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Dec. 2020, pp. 1877–1901.
[47] Yinhan Liu, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," in *International Conference on Learning Representations (ICLR)*, M. O. myleott@fb. com , N. G. J. D. M. J. D. C. O. L. M. L. L. Z. V. S. Yinhan Liu, Ed.,
[48] William Fedus, Barret Zoph∗, and Noam Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," *Journal of Machine Learning Research*, vol. 23, pp. 1–39, 2022.
[49] Edward Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," 2021.
[50] Tri Dao, Daniel Y. Fu, Stefano Ermon, C. R. Atri Rudra, and Christopher Ré, "FLASHATTENTION: fast and memory-efficient exact attention with IO-awareness," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Nov. 2022, pp. 16344–16359.
[51] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism," *ArXiv, abs/1909.08053.*, Sep. 2019.