

DEPARTMENT OF ELECTRICAL & ELECTRONICS ENGINEERING

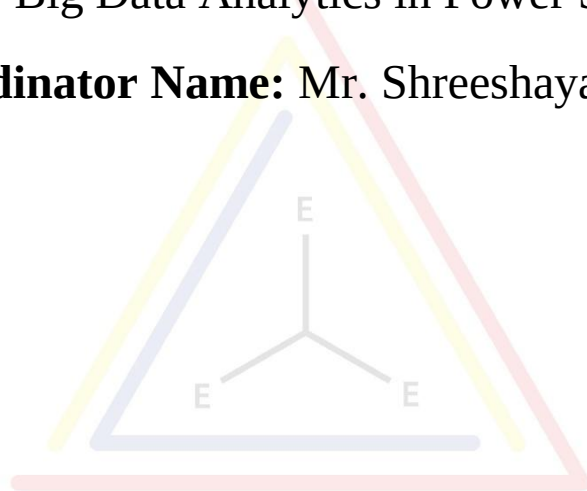
NOTES

Academic Year: 2026-27

Course Code : BEE714D

Course Title : Big Data Analytics in Power Systems

Course Co-ordinator Name: Mr. Shreeshayana R



Department of EEE
Emitting Elite Energy

DEPARTMENT OF ELECTRICAL & ELECTRONICS ENGINEERING

Vision of the Institute Development of academically excellent, culturally vibrant, socially responsible and globally competent human resources. Mission of the Institute • To keep pace with advancements in knowledge and make the students competitive and capable at the global level. • To create an environment for the students to acquire the right physical, intellectual, emotional and moral foundations and shine as torch-bearers of tomorrow's society. • To strive to attain ever-higher benchmarks of educational excellence	Vision of the Department To create Electrical and Electronics Engineers who excel to be technically competent and fulfil the cultural and social aspirations of the society. Mission of the Department • To provide knowledge to students that builds a strong foundation in the basic principles of electrical engineering, problem solving abilities, analytical skills, soft skills and communication skills for their overall development. • To offer outcome based technical education. • To encourage faculty in training & development and to offer consultancy through research & industry interaction.
Program Educational Objectives (PEO's) • PEO1: To produce competent and ethical Electrical and Electronics Engineers who will exhibit the necessary technical and managerial skills to perform their duties in society. • PEO2: To make graduates continuously acquire and enhance their technical and socio-economic skills. • PEO3: To aspire graduates on R&D activities leading to offering solutions and excel in various career paths. • PEO4: To produce quality engineers who have the capability to work in teams and contribute to real time projects.	Program Specific Outcomes (PSO's) • PSO1: Apply the concepts of Electrical & Electronics Engineering to evaluate the performance of power systems and also to control industrial drives using power electronics. • PSO2: Demonstrate the concepts of process control for Industrial Automation, design models for environmental and social concerns and also exhibit continuous self- learning.

DEPARTMENT OF ELECTRICAL & ELECTRONICS ENGINEERING

Program Outcomes (PO's)

PO1: Engineering Knowledge: Apply knowledge of mathematics, natural science, computing, engineering fundamentals and an engineering specialization as specified in WK1 to WK4 respectively to develop to the solution of complex engineering problems.

PO2: Problem Analysis: Identify, formulate, review research literature and analyze complex engineering problems reaching substantiated conclusions with consideration for sustainable development. (WK1 to WK4)

PO3: Design/Development of Solutions: Design creative solutions for complex engineering problems and design/develop systems/components/processes to meet identified needs with consideration for the public health and safety, whole-life cost, net zero carbon, culture, society and environment as required. (WK5)

PO4: Conduct Investigations of Complex Problems: Conduct investigations of complex engineering problems using research-based knowledge including design of experiments, modelling, analysis & interpretation of data to provide valid conclusions. (WK8).

PO5: Engineering Tool Usage: Create, select and apply appropriate techniques, resources and modern engineering & IT tools, including prediction and modelling recognizing their limitations to solve complex engineering problems. (WK2 and WK6)

PO6: The Engineer and The World: Analyze and evaluate societal and environmental aspects while solving complex engineering problems for its impact on sustainability with reference to economy, health, safety, legal framework, culture and environment. (WK1, WK5, and WK7).

PO7: Ethics: Apply ethical principles and commit to professional ethics, human values, diversity and inclusion; adhere to national & international laws. (WK9)

PO8: Individual and Collaborative Team work: Function effectively as an individual, and as a member or leader in diverse/multi-disciplinary teams.

PO9: Communication: Communicate effectively and inclusively within the engineering community and society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations considering cultural, language, and learning differences

PO10: Project Management and Finance: Apply knowledge and understanding of engineering management principles and economic decision-making and apply these to one's own work, as a member and leader in a team, and to manage projects and in multidisciplinary environments.

PO11: Life-Long Learning: Recognize the need for, and have the preparation and ability for i) independent and life-long learning ii) adaptability to new and emerging technologies and iii) critical thinking in the broadest context of technological change. (WK8)

MODULE 1: Introduction

Structure

- 1.1 Introduction: Big Data,
- 1.2 Future Power Systems.
- 1.3 Big Data Application and Analytics in a Large - Scale Power System: Introduction
- 1.4 General Applications of Big Data
- 1.5 Algorithms for Processing Big Data
- 1.6 Application of Big Data in Power Systems.

Objectives of Course

Learning Objectives

1. To define big data and to explain big data application and analytics to power systems.
2. To explain the role of big data in smart grid communications and optimization of big data in electric power systems.
3. To explain security methods for the infrastructure communication and data mining methods for theft detection in power systems.
4. To explain the application of unit commitment method in the control of smart grid.
5. To explain protection algorithm for transformer based on data pattern recognition

1.1 Introduction: Big Data

- **Definition – Big Data** refers to extremely **large datasets** that cannot be processed efficiently using traditional data processing techniques.
 - **Characteristics (5 Vs)** – Volume (size), Velocity (speed), Variety (different types), Veracity (quality/reliability), and Value (usefulness).
 - **Sources** – Generated from social media, IoT devices, sensors, business transactions, healthcare, and industrial machines.
 - **Challenges** – Data storage, processing, real-time analysis, and maintaining data security.
 - **Technologies Used** – Hadoop, Spark, NoSQL databases, cloud computing platforms.
 - **Applications** – Power systems, healthcare, finance, e-commerce, weather forecasting, and smart cities.
-

1.2 Future Power Systems.

Characteristics of Future Power Systems

- Increased **decentralization** with microgrids and nanogrids.
- **Expanded communication and monitoring** capabilities.
- **Wider variety of energy sources**, including more renewables.

Research Thrusts in Future Power Systems

- Expansion of the **Smart Grid** to improve monitoring and control.
- **Internet of Things (IoT)** integration for device-level communication.
- Increased **renewable energy penetration**.
- **Microgrid and nanogrid deployment** for local resiliency.

Smart Grid Implications

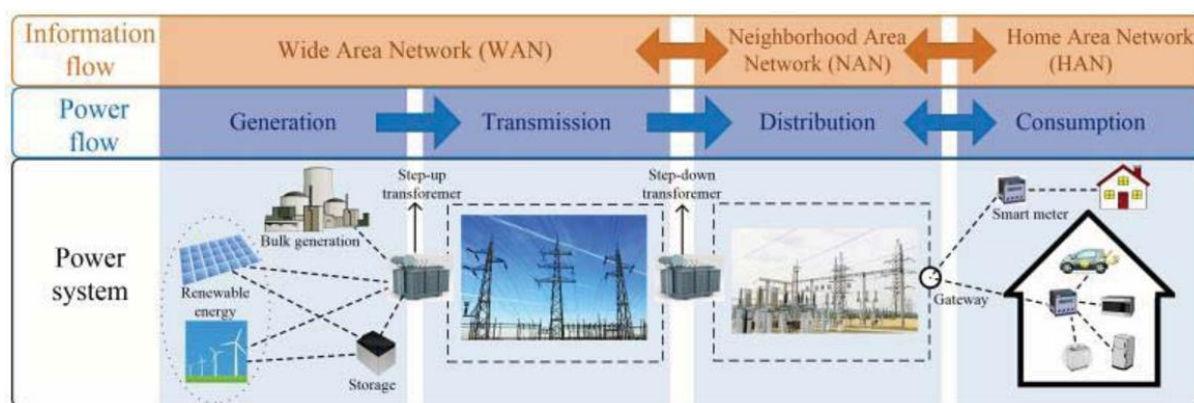
- Enables **data collection** for theft detection and grid optimization.
- Generates **big data challenges** due to the volume and velocity of collected data

IoT-enabled Power Grid

- Allows **communication with all devices** in the network.
- Provides **real-time monitoring of critical infrastructure (CI)**.
- Introduces **big data and cybersecurity challenges**.

Microgrids and Renewables

- Offer **local resiliency** but create **uncertainty for larger grid planning**.
- Renewable energy sources add **availability and reliability challenges**.



1.3 Big Data Application and Analytics in a Large - Scale Power System: Introduction

Big Data Analysis Methods

- Relies heavily on **machine learning techniques**.
- Machine learning is associated with **pattern recognition, statistics, and data mining**.

Emergence of Advanced Techniques

- **Deep learning** (large-scale neural networks) is increasingly used.
- Capable of handling the **size and complexity** of big data.

Applications in Power Systems

- Initially successful in **image recognition** tasks.
- Now being **applied to big data analysis in power systems** for enhanced insights and decision-making.

Big Data Analysis Methods in Power Systems

1. Role of Machine Learning (ML)

1. ML algorithms are essential for handling large-scale, complex datasets generated by modern power grids.
2. ML is associated with **pattern recognition, statistics, and data mining**, enabling automatic detection of patterns and anomalies in grid operations.

2. Applications in Power Systems

- **Load Forecasting:** Predicting electricity demand for better grid management.
- **Fault Detection and Diagnosis:** Recognizing abnormal patterns in sensor data to prevent outages.
- **Power Theft Detection:** Identifying irregular consumption patterns from smart meter data.
- **Renewable Energy Integration:** Forecasting solar and wind power output to optimize generation and storage.

3. Advanced Methods

1. **Deep Learning (DL):** Large-scale neural networks analyze high-dimensional, real-time data such as PMU (Phasor Measurement Unit) streams.
2. **Predictive Analytics:** Uses historical and live data to make proactive grid management decisions.

3. **Data Mining:** Extracts hidden trends from vast power system databases for planning and optimization.

1.4 General Applications of Big Data

Beyond Syllabus

Sector	Key Use Cases
Healthcare	Personalized treatment, predictive risk, diagnostics
Marketing	Targeting, personalization, campaign optimization
Transportation	Smart traffic, route optimization, congestion control
Finance	Fraud detection, risk analytics, compliance
Retail	Recommendations, dynamic pricing, inventory management
Government	Public safety, fraud control, social services
Manufacturing	Predictive maintenance, efficiency in operations
Education	Student performance analysis, learning insights
Agriculture	Crop monitoring, satellite insights, yield boost
Cybersecurity	Behavioral threat detection, anomaly identification

Key Applications of Big Data

- **Healthcare:** Enables personalized medicine, predictive analytics for clinical risk, automated reporting, and improved diagnostics using vast and varied medical data.
- **Marketing & Advertising:** Supports precise customer targeting and personalized messaging by analyzing behavior and preferences at scale.
- **Transportation & Smart Cities:** Powers intelligent traffic systems, real-time route planning, congestion prediction, and urban computing solutions like GPS-based path estimation
- **Finance & Banking:** Used for fraud detection, credit scoring, risk management, portfolio optimization, and regulatory compliance
- **Retail & E-commerce:** Enhances customer experience through recommendation systems, inventory optimization, dynamic pricing, and demand forecasting.
- **Government & Public Sector:** Facilitates fraud detection, public health tracking, environmental protection, and efficient public service delivery.
- **Manufacturing & Industrial:** Enables predictive maintenance, smart factory operations, real-time analytics, and supply chain visibility.
- **Education & Learning Analytics:** Helps track student engagement, performance trends, and improve teaching effectiveness
- **Climate, Agriculture & Earth Science:** Supports satellite data analysis for precision agriculture, yield forecasting, and climate trend monitoring.
- **Cybersecurity & Behavioral Analytics:** Detects anomalies, predicts potential fraud or threats, and analyzes user behavior to secure systems.

As per Text Book:**Introduction**

- Big Data grew rapidly due to the widespread use of the Internet.
- The Internet acts as:
 - A broadcasting medium
 - A tool for information sharing
 - A platform for collaboration regardless of location
- Massive amounts of data are generated daily.
 - Example: Google processes 20+ petabytes/day of user data.
- Data sources:
 - Web browsing
 - E-commerce purchases
 - Bank/credit card transactions
 - Social networks
- The Big Data explosion impacts all businesses and industries.
- Healthcare and social networking are key examples of Big Data's wide-scale influence.

Applications**1. Health Care – Sources of Big Data****Eight major sources (Gartner 2016):**

1. Physicians' free-text notes
2. Patient-generated health data (PGHD)
3. Genomics
4. Physiological monitoring data
5. Publicly available data
6. Credit card & purchasing data
7. Social media data
8. Medical imaging data

Data Volume & Structure:

- Genomics: ~100 GB per patient
- Physicians' notes: Terabytes (unstructured text)
- Medical imaging: Petabytes
- Data ranges from unstructured (notes) → structured formats (genomics).

Health Care Analytics – Importance & Challenges

- Healthcare analytics is vital due to:
 - Growing demand for information
 - Large, diverse datasets
 - Increased competition
 - Complex regulations
- **Innovations:**
 - Precision medicine
 - Value-based care
 - Population health management
- **Value-based care:**

- Relies on strong data & analytics
- Requires heavy investment in analytic infrastructure
- Stakeholders: providers, systems, insurers, pharma, life sciences
- **Key Needs:**
 - Improve quality of information
 - Establish analytics programs
 - Data governance & IT platforms
 - Better information management
- **Sources:** electronic health records, claims, wearable devices, social media, patients.
- **Analytics can:**
 - Detect patterns in information
 - Deliver actionable insights
 - Enable predictive/self-learning systems (predict, infer, suggest alternatives).
- **Benefits:**
 - Reduce costs & improve quality
 - Identify and treat at-risk populations
 - Connect better with patients/consumers
 - Evaluate health interventions' impact
- **Challenges in healthcare business:**
 - Measure & improve clinical performance
 - Ensure clinical quality of care
 - Assess outcomes (mortality, infections, survival, treatment paths)
- **Raw data problem:**
 - Data are siloed → must be transformed into patient-centered structures
 - Data should be accessible to patients & stakeholders
 - Real-time predictive models help solve clinical & operational issues
- **Key goal of Big Data in healthcare:**
 - Identify which data sources improve analytics → support clinical decision-making & healthcare improvements.

2. Social Networking

1. Social networking depends heavily on the Internet.
2. Growth in **size, complexity, and unstructured data** has been explosive.
3. Research uses big data in social networking through **observations, experiments, and simulations**.
4. Companies like **Facebook, Google, Microsoft** collect vast data from calls, texts, clicks, etc.
5. This data is often kept private for **competitive reasons** and **user privacy protection**.

Handling Big Data

- Big data management relies on two key aspects:
 1. **Hardware capability** (storage, bandwidth, processing power).
 2. **Applications/algorithms** (social networking, machine learning, data mining, cloud computing).
- Hardware improvements allow **higher capacity at lower cost**.
- Algorithms enhance scalability, analytics, and predictive capabilities.

- Example: **Cost of 1 GB storage** dropped from **\$300,000 (1981)** → **\$1,000 (1994)** → just a few cents today.

1.5 Algorithms for Processing Big Data

Algorithms for Processing Big Data

- Algorithms for big data analysis include machine learning and deep learning.
- Deep learning is a subset of machine learning but uses advanced methods and architectures.

Machine Learning and Deep Learning Generalities

- **Machine Learning (ML):** Uses training data to classify unseen data.
- **Algorithms define weights for models and update them until convergence.**
- **Two types:**
 1. **Supervised Learning:** Uses labeled data (input + output known). Learns mapping rules.
 2. **Unsupervised Learning:** Uses unlabeled data. Aims to find structure/patterns in data.

Machine Learning

- ML handles large and complex datasets effectively.
- Provides classification and prediction algorithms for quick results.
- Useful in **real-time decision making**, e.g.:
 1. Power market management.
 2. System stability analysis.
 3. Electricity market predictions.
 4. Customer load forecasting.
 5. Power system state estimation.

1.5.1 Artificial Neural Network (ANN) Model

- Neural networks are effective at understanding complex data patterns.
- They can detect patterns too complex for humans or traditional computer methods.
- ANN consists of layers of interconnected processing nodes (neurons).
- The Multi-layer Perceptron (MLP) is the most commonly used ANN architecture.
- **A typical ANN has three layers:**
 1. **Input Layer** – receives data from the environment.
 2. **Hidden Layer** – processes data and connects input to output.
 3. **Output Layer** – delivers the final response or classification.
- Neurons are represented as circular nodes; connections (with weights) represent data flow between them.
- Weighted connections link the output of one neuron to the input of another.
- Hidden layers help extract complex information for classification.

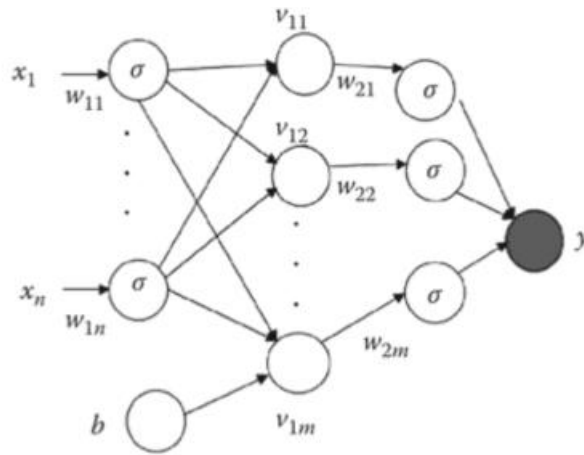


Fig: Feedforward Neural Network

Steps Describing ANN Neuron and Learning Process:

Step-1: Input and Output Definition

1. Input and Output Definition

- The **input layer** is a vector:

$$X = \{x_1, x_2, \dots, x_n\}$$
- The **output** is denoted as:

$$y$$

Step-2: Neuron Computation in Hidden Layer

- A neuron k in the j^{th} hidden layer is calculated as:

$$v_{jk} = \sigma \left(\sum_{i=0}^n (w_{ji}x_i + b) \right)$$

- Where:
 - $\sigma(\cdot)$: Activation function
 - $w_j = \{w_{j1}, w_{j2}, \dots, w_{jn}\}$: Weight vector
 - b : Bias value

Step-3: Goal of ANN Learning

- The objective is to find a weight set $W = \{w_1, w_2\}$ that minimizes the error.
- Error is typically calculated using the **sum of squared errors**.

Step-4: Error (Cost) Function

- The cost of the error is given by:

$$E = [y - f(W, X)]^2$$

- Where:
 - $f(W, X)$: Output predicted by the ANN
 - y : Actual (desired) output

1.5.2 Support Vector Machine (SVM)

- **SVM** is an algorithm like ANN used for classification and regression tasks.
- It identifies a **function mapping** that splits input data into separate classes.
- The goal is to:
 - Map data into a **new space** where it becomes **linearly separable**.
 - Perform **linear classification** in this new space.

Step-1: Regression SVM Problem Statement

- **Training dataset:**
 $\{(X_i, y_i)\}$ for $i = 1, 2, \dots, n$
 Where:
 - X_i : input vector
 - y_i : output vector
- **Objective:**
 Find a function $y(X)$ that relates input features to outputs and predicts the output for new inputs.

Step-2: Feature Mapping

- Input X is mapped to **m-dimensional feature space** using:
 $g_j(x), \quad j = 1, 2, \dots, m$

Step-3: SVM Output Function

- The SVM function is given by:

$$y(x, \omega) = \sum_{j=1}^m \omega_j g_j(x) + b$$

Where:

- ω_j : weights
- $g_j(x)$: feature mapping functions
- b : bias term

Step-4: ϵ -Insensitive Loss Function

- The loss function ignores errors less than a threshold ε :

$$e_{\varepsilon}(r, y(x, \omega)) = \begin{cases} 0 & \text{if } |r - y(x, \omega)| \leq \varepsilon \\ |r - y(x, \omega)| - \varepsilon & \text{otherwise} \end{cases}$$

Step-5: Optimization Objective

- The goal of SVM regression is to **minimize** both:
 - The model complexity (via weight norm $\|\omega\|^2$)
 - The training error (via slack variables ξ_i, ξ_i^*)

$$\min \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*)$$

Step-6: Subject to Constraints

$$\begin{cases} r_i - y(x_i, \omega) \leq \varepsilon + \xi_i^* \\ y(x_i, \omega) - r_i \leq \varepsilon + \xi_i \\ \xi_i, \xi_i^* \geq 0, \quad i = 1, \dots, n \end{cases}$$

- ξ_i, ξ_i^* : Slack variables representing **lower and upper training errors**

Step-7: Dual Formulation of SVM Regression

1. Dual Form Equation

- The dual form of the SVM regression function is:

$$y(x) = -\frac{1}{2} \sum_t \sum_s (\alpha_t^+ - \alpha_t^-)(\alpha_s^+ - \alpha_s^-) K(x_t, x_s) - \varepsilon \sum_s (\alpha_s^+ + \alpha_s^-) + \sum_s r_s (\alpha_s^+ - \alpha_s^-)$$

2. Constraints

$$\begin{cases} 0 \leq \alpha_t^+ \leq C \\ 0 \leq \alpha_t^- \leq C \\ \sum_t (\alpha_t^+ - \alpha_t^-) = 0 \end{cases}$$

3. Kernel Function

- The kernel function $K(x, x_i)$ is used to map inputs to a higher-dimensional space:

$$K(x, x_i) = \sum_{j=1}^m g_j(x) g_j(x_i)$$

4. Role of Parameter C

- C controls the **trade-off** between:
 - Model complexity (large C $\rightarrow$ complex model)
 - Tolerance to error (small C $\rightarrow$ more error allowed)
- Higher C reduces training error but may increase overfitting.

5. Use of Kernel Functions

- The function $K(x, x_i)$ can be estimated using **kernel functions**.
- A commonly used kernel in SVMs is the **Radial Basis Function (RBF)** kernel.

Step-8: RBF

RBF (Radial Basis Function) Kernel

- The RBF kernel is defined as:

$$K(x_i, x) = \exp\left(-\frac{\|x - x_i\|^2}{2p^2}\right)$$

- This kernel measures similarity between inputs, where p controls the kernel width (spread of the function).

Step-9: Final SVM Regression Function

- After solving the dual problem (Eq. 2.6), the coefficients α_i^+ , α_i^- are obtained.
- The final prediction function becomes:

$$y(x) = \sum_{i=1}^{n_{SV}} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

- This is a **weighted sum of support vectors**.

Example-1:

Imagine you want to predict the price of a house based on its size.

- **Step 1: Data** — You have house sizes (input) and their prices (output).
- **Step 2: Map Features** — Sometimes, the relationship isn't simple, so you transform the size data into a new form (feature mapping) that makes it easier to predict price linearly.
- **Step 3: Find the best function** — SVM tries to find a function (a line or curve) that predicts house prices as closely as possible.
- **Step 4: ϵ -Insensitive Loss** — SVM ignores small errors within a margin (ϵ) — small differences between predicted and actual prices are acceptable.
- **Step 5: Optimization** — It balances between fitting the data well and keeping the function smooth to avoid overfitting.
- **Step 6: Kernel Trick (e.g., RBF)** — If data isn't linearly predictable, it uses a kernel (like RBF) to map data into a higher-dimensional space where linear prediction works.
- **Step 7: Final Model** — The model predicts house prices as a combination of important "support vectors" (key training points) weighted properly.

In short:

SVM regression finds a function that **predicts outputs from inputs**, allowing some small errors, while **balancing accuracy and simplicity** — often transforming data to make predictions easier.

Example-2:

Imagine you want to classify emails as Spam or Not Spam based on the number of certain keywords.

- **Step 1: Data** — You have emails with features like the count of suspicious words, and labels (Spam or Not Spam).
- **Step 2: Map Features** — Sometimes the data can't be separated by a simple line, so features are transformed to a new space.
- **Step 3: Find the best boundary** — SVM finds the line (or hyperplane) that best separates Spam emails from Not Spam emails.
- **Step 4: Margin Maximization** — It chooses the boundary that keeps the biggest gap (margin) between the two groups.
- **Step 5: Handle errors** — Some emails might be misclassified to keep the margin large enough (soft margin).
- **Step 6: Kernel Trick** — If emails aren't separable in original features, SVM uses a kernel (like RBF) to separate them in a higher-dimensional space.
- **Step 7: Final Model** — The decision boundary is formed mainly by a few important emails called "support vectors."

1.5.3 Decision-Tree Classifier

1. Overview

- A Decision Tree is used for multi-stage decision-making.
- It follows a recursive top-down approach.

2. Structure

- **Root node:** The starting point of the tree (no incoming edges).
- **Internal nodes:** Perform recursive partitioning of input space using decision rules.
- **Leaf nodes:** Terminal nodes representing the final class labels (no outgoing edges).

3. Working

- The input space is partitioned into **two or more subclasses** at each internal node.
- **Recursion continues** until all data points are classified into appropriate leaf nodes.
- Each **leaf node** is assigned one class that best fits the data reaching that point.

CART Algorithm

- **CART:** Binary decision tree algorithm for **classification** and **regression**.
- **Splitting:** Chooses features that maximize **purity** (e.g., Gini Index, MSE).
- **Classification:** Uses **Gini Impurity** or **Entropy**.
- **Regression:** Uses **variance reduction** or **Mean Squared Error (MSE)**.
- **Tree growth:** Built **recursively**.
- **Pruning:** Applied to reduce **overfitting**.

Step-1: Gini Index**1. Purpose:**

- Measures the **impurity** of a node in the **CART algorithm**.

2. Definition:

- Applied to node i , which processes dataset S_i from its parent node.

3. Formula:

$$\text{Gini}(S_i) = 1 - \sum_{k=1}^K (F_{k,i})^2$$

4. Where:

- $F_{k,i}$: Fraction of data in S_i belonging to class k
- K : Total number of classes

5. Interpretation:

- Minimum Gini (0)**: All data belongs to **one class** (pure node).
- Maximum Gini**: Data is **evenly distributed** among all classes (highest impurity).

Step-2: Gini Gain**1. Attribute Selection:**

- For each attribute $j \in \{1, 2, \dots, N\}$, its values A^j are split into two disjoint subsets:
 - A_L^j (left)
 - A_R^j (right)

2. Subset Properties:

- $A_L^j \cup A_R^j = A^j$ (they are **complementary** subsets)

3. Partitioning:

- All possible partitions from set P_i divide dataset S_i at node i into two:
 - Left subset: $L_i(A_L^j, A_R^j)$
 - Right subset: $R_i(A_L^j, A_R^j)$

4. Fraction of Data:

- Let:
 - $F_L(A_L^j, A_R^j)$: Fraction in left subset
 - $F_R(A_L^j, A_R^j)$: Fraction in right subset

5. Constraint:

- The sum of fractions equals 1:

$$F_L(A_L^j, A_R^j) + F_R(A_L^j, A_R^j) = 1$$

Step-3: Gini Gain Calculation**1. Fraction of Data per Class:**

- For class $k \in \{1, 2, \dots, K\}$, fractions of data in subsets are:
 - $F_{L,i,k}(A_L^j, A_R^j)$ for left subset
 - $F_{R,i,k}(A_L^j, A_R^j)$ for right subset

2. Weighted Gini Index for partition sets A_L^j and A_R^j :

$$\text{Weighted_Gini}(S_i, A_L^j, A_R^j) = F_{L,i}(A_L^j, A_R^j) \cdot \text{Gini}(L_i(A_L^j, A_R^j)) + F_{R,i}(A_L^j, A_R^j) \cdot \text{Gini}(R_i(A_L^j, A_R^j))$$

3. Gini for Left and Right subsets:

$$\text{Gini}(L_i(A_L^j, A_R^j)) = 1 - \sum_{k=1}^K \left(F_{L,i,k}(A_L^j, A_R^j) \right)^2$$

$$\text{Gini}(R_i(A_L^j, A_R^j)) = 1 - \sum_{k=1}^K \left(F_{R,i,k}(A_L^j, A_R^j) \right)^2$$

4. Gini Gain calculation:

$$G(S_i, A_L^j, A_R^j) = \text{Gini}(S_i) - \text{Weighted_Gini}(S_i, A_L^j, A_R^j)$$

Example: Classify Animals as Cat or Dog based on Tail Length**Animal Tail Length Label**

1	Short	Cat
2	Short	Cat
3	Long	Dog
4	Long	Dog
5	Short	Dog

Step 1: Calculate Gini Index for whole data

- Total animals = 5
- Cats = 2, Dogs = 3

$$p_{Cat} = \frac{2}{5} = 0.4, \quad p_{Dog} = \frac{3}{5} = 0.6$$

$$Gini = 1 - (0.4^2 + 0.6^2) = 1 - (0.16 + 0.36) = 0.48$$

Step 2: Split by Tail Length (Short vs Long)

- Short Tail:** Animals 1, 2, 5
Cats = 2, Dogs = 1

$$p_{Cat} = \frac{2}{3} \approx 0.67, \quad p_{Dog} = \frac{1}{3} \approx 0.33$$

$$Gini = 1 - (0.67^2 + 0.33^2) = 1 - (0.45 + 0.11) = 0.44$$

- Long Tail:** Animals 3, 4
Cats = 0, Dogs = 2

$$p_{Dog} = 1, \quad Gini = 0 \text{ (pure node)}$$

Step 3: Calculate weighted Gini after split

$$\text{Weighted Gini} = \frac{3}{5} \times 0.44 + \frac{2}{5} \times 0 = 0.264$$

Step 4: Calculate Gini Gain

$$\text{GiniGain} = \text{Original Gini} - \text{Weighted Gini} = 0.48 - 0.264 = 0.216$$

Conclusion:

- Splitting by tail length reduces impurity by 0.216 (improves classification).
- Decision Tree will split first on **Tail Length**.

Simple takeaway:

Gini Index measures impurity, Gini Gain shows how well a split separates classes, and the tree picks splits with highest gain.

1.5.4 Deep Learning Models

1. Deep learning analyses big data by using complex data representations and abstractions.
2. It is effective for decision-making, information retrieval, and classification, especially with unsupervised data.
3. Deep learning models use stacked layers where each layer performs nonlinear transformations.
4. Data flows hierarchically through these layers, automating feature extraction at each step.
5. More layers enable deeper, more complex feature extraction and richer data representation.
6. The final model output is a highly nonlinear function of the original input data.
7. Deep learning works with both supervised (labelled) and unsupervised (unlabelled) data.
8. **Convolutional Neural Networks (CNNs)** are a type of deep learning model with convolutional and pooling layers.
9. Convolutional layers apply learnable filters to extract features; pooling layers reduce data size via averaging or max-pooling.
10. Lower CNN layers capture simple features, while higher layers capture more abstract, complex features, with layer depth depending on problem complexity.

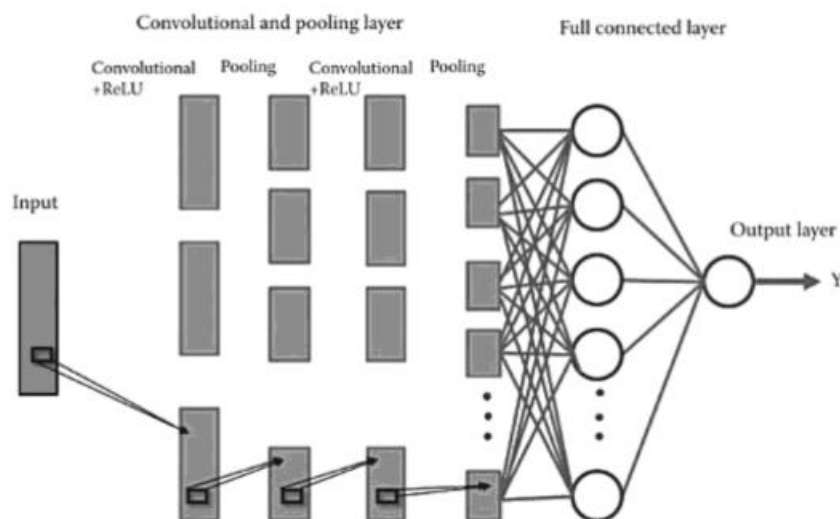


Fig: Scheme of Feedforward CNN

Example: Recognizing Handwritten Digits (0-9)

1. **Input:** An image of a handwritten digit (like '7').
2. **Convolutional Layers:** The CNN applies filters to detect simple features such as edges and curves in the image.
3. **Pooling Layers:** These reduce the image size while keeping important features, making the model faster and more efficient.
4. **Deeper Layers:** Later layers combine simple features to detect more complex patterns like loops or intersections.
5. **Output Layer:** Finally, the model predicts which digit (0-9) the image represents.

Summary:

CNNs automatically learn to extract features from images layer by layer, improving accuracy in recognizing patterns like handwritten digits.

Example: Diagnosing Pneumonia from Chest X-Ray Images

1. **Input:** Chest X-ray images of patients' lungs.
2. **Convolutional Layers:** The CNN applies filters to detect edges, textures, and patterns indicating lung abnormalities.
3. **Pooling Layers:** These reduce the size of the image data while preserving important features like spots or shadows.
4. **Deeper Layers:** Later layers combine these basic features to identify complex signs of pneumonia, such as specific shapes or textures in lung tissue.
5. **Output Layer:** The model outputs a prediction—whether the patient has pneumonia or not.

Why CNN works well here:

- Automatically extracts important features from raw images.
- Handles the complexity and variability in medical images.
- Reduces the need for manual feature engineering by experts.

1.6 Application of Big Data in Power Systems.

1. Big Data in Smart Grid Networks

- **Smart Grid** = Digital upgrade of traditional electricity grids.
- Equipped with **smart meters, IoT sensors, communication networks, and control systems**.
- Generates real-time data on **voltage, current, frequency, consumption, and system health**.

Applications:

1. **Demand Forecasting** – Predicting electricity demand patterns (hourly/daily/seasonal).
2. **Fault Detection** – Identifying outages or abnormal conditions quickly.
3. **Energy Theft Detection** – Data analytics can flag unusual consumption.
4. **Grid Optimization** – Deciding which lines/generators to use for minimum loss.

Example: In India, smart meters in Delhi DISCOMs provide data every 15 minutes, enabling quick detection of theft and billing errors.

2. Phasor Measurement Units (PMU)

- **PMUs** measure voltage and current phasors (magnitude + phase angle) at very high speed (30–60 samples/second).
- Data is time-synchronized using **GPS signals**, enabling **wide-area monitoring**.

Applications of Big Data in PMU:

1. **Early Warning Systems** – Detects oscillations that may cause blackouts.
2. **Grid Stability Analysis** – Helps prevent cascading failures.
3. **State Estimation** – Improves accuracy of real-time grid modeling.

Example: After the 2003 US–Canada blackout, North America deployed thousands of PMUs to provide wide-area situational awareness.

3. Renewable Energy

- Solar and wind are **intermittent** → variability creates challenges for grid operators.
- Big Data analytics uses **weather data, satellite images, and historical generation data** for forecasting.

Applications:

1. **Forecasting Solar/Wind Output** – Improves scheduling of backup power.
2. **Integration into Grid** – Avoids instability by predicting fluctuations.
3. **Market Participation** – Renewable producers can sell power more accurately in day-ahead markets.

Example: Tamil Nadu (India) has one of the largest wind farms. Forecasting tools help balance renewable energy with thermal generation.

4. CIM as Information Standard for Big Data Analytics

- **CIM (Common Information Model)** is an international standard (IEC 61970/61968).
- It provides a **common data model** for exchanging power system information between different devices, utilities, and vendors.

Why Important?

- Power system data comes from many sources → different formats.
- CIM ensures **interoperability** and allows Big Data platforms to integrate datasets smoothly.

Example: European utilities use CIM to integrate data from multiple renewable sources and grid operators.

Big Data Problem in Power System Modeling

Power system optimization problems are very **large-scale** and complex. Big Data creates both **opportunities** and **challenges**.

1. Security-Constrained Unit Commitment (SCUC)

- **Unit Commitment** = deciding which power plants to turn ON/OFF and at what output level.
- SCUC adds **security constraints** → ensures grid remains stable even if a generator/line fails.

Big Data Role:

- Faster simulations with real-time data.
- Uses machine learning to predict demand, prices, and failures.

Example: In Indian Energy Exchange (IEX), SCUC ensures backup power plants are available during peak load.

2. Decomposition Methods to Handle Big Data

- Large power system optimization problems are too complex for one computer.
- **Decomposition** splits them into smaller problems (e.g., regional subproblems).
- Each subproblem is solved separately, then results are combined.

Techniques:

- Benders Decomposition
- Lagrangian Relaxation

Benefit: Faster solutions for large systems (like all-India grid).

3. Firm Transmission Right (FTR) Problems

- **FTRs** = financial contracts that protect utilities from congestion charges.
- When transmission lines are congested, FTR holders get compensation.

Big Data Role:

- Predicts congestion patterns using historical + real-time data.
- Optimizes allocation of FTRs in electricity markets.

Example: In PJM (US electricity market), billions of dollars in FTRs are traded annually, supported by Big Data analytics.

4. Time-Constrained Economic Dispatch

- **Economic Dispatch** = determining generation levels to meet demand at minimum cost.
- Must be solved **within seconds** in real-time markets.

Big Data Role:

- Uses high-performance computing + data analytics.
- Improves accuracy of cost minimization under tight time limits.

Example: Indian real-time market (RTM) introduced in 2020 requires quick dispatch decisions every 15 minutes, enabled by advanced data platforms.

Outcomes

At the end of the module, students will be able to:

CO-1: Interpret the role of big data and machine-learning methods applicable to power systems and in particular to Smart Grid communications. [L2]

TEXT BOOKS:

Big Data Analytics in Future Power Systems, Ahmed F. Zobaa and Trevor J. Bihl, CRC Press 2019. 2019.

Reference Books/ Link

1. **Big Data Analytics for Power Systems** – [Big Data Analytics in Power Systems](#)
2. **Application of Big-Data Analytics in Power System Protection**-[Lec-37: Application of Big-Data Analytics in Power System Protection](#)

MODULE 2: Role of Big Data in Smart Grid Communications:

Structure

- 2.1 Introduction,
- 2.2 The Grid Modernization,
- 2.3 The Grid Interconnection with the Internet of Things,
- 2.4 Data Traffic Pattern in a Smart Grid Environment,
- 2.5 The Massive Flow of Information in a Smart Scenario,
- 2.6 The Volume of Generated Data in a Smart Distribution System: A Case of Study.

Objectives of Course

Learning Objectives

To explain the role of big data in smart grid communications and optimization of big data in electric power systems.

2.1 Introduction

- Smart Grid as a Sensor Network – Consists of many connected smart devices.
- Rising Data Flow – Growth of devices and operational needs increases information exchange.
- Big Data Role – Provides efficient management of large data volumes.
- Benefits – Helps utilities understand customer behavior, demand, weather, outages, and failures.
- Robust Methodologies – Needed to make the grid smarter and more reliable.
- Key Concern – Quantifying and utilizing data from devices effectively.
- Chapter Aim – To characterize and evaluate data growth in smart grid communication networks.
- **Example – Active distribution system shows massive data generation for monitoring and control.**

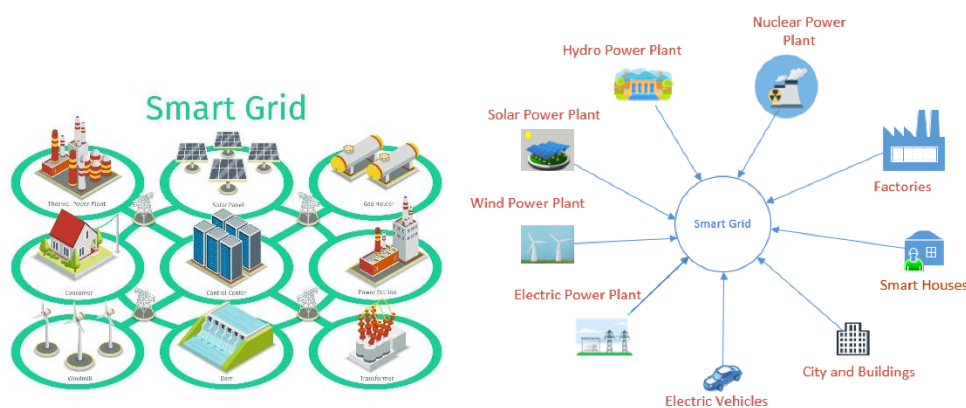


Fig-1: Smart Grid

2.2 The Grid Modernization,

- **Dynamic Nature** – Power generation variations directly affect customer supply → requires real-time monitoring and operation.
- **Smart Grid** – Integrates ICT with electrical system to improve reliability, robustness, safety, reduce losses and failures (especially at distribution level where 90% failures occur).
- **Distributed Generation**
 (Solar/Wind) – Higher penetration increases need for monitoring and efficient info exchange among all agents.

- **Role of IEDs** – Real-time automation through data from user profiles, PMUs, distributed resources, AMI, sensors, actuators, breakers, capacitor banks, and utilities.
- **Advanced Metering Infrastructure (AMI)** – First step in modernization; enables two-way communication between users and utilities, supports demand response, pricing, state estimation, and real-time supply-demand balance.
- **Phasor Measurement Units (PMUs)** – Applied in protection, state estimation, voltage/frequency control, renewable integration, and islanding monitoring.
- **Massive Data Generation** – Intelligent devices and sensors create huge volumes of information requiring proper management.
- **Big Data as Solution** – Provides tools to handle and analyze large datasets, improving decision-making, grid robustness, and profitability.
- **Key Requirement** – Quantify which devices generate data and how much, to effectively apply Big Data technologies.

2.3 The Grid Interconnection with the Internet of Things

- **Intelligent Communication Network** – Must be high-performance, reliable, robust, flexible → handles data collection, routing, monitoring, and management.
- **Scale of Communication** – Thousands of devices send data to substations and control centers with minimal human interaction; each distribution system has hundreds–thousands of smart meters, PMUs, and IEDs.
- **IoT in Smart Grids** – Enhances flexibility, scalability, and automation → forms a cyber-physical system integrating power and communication networks.

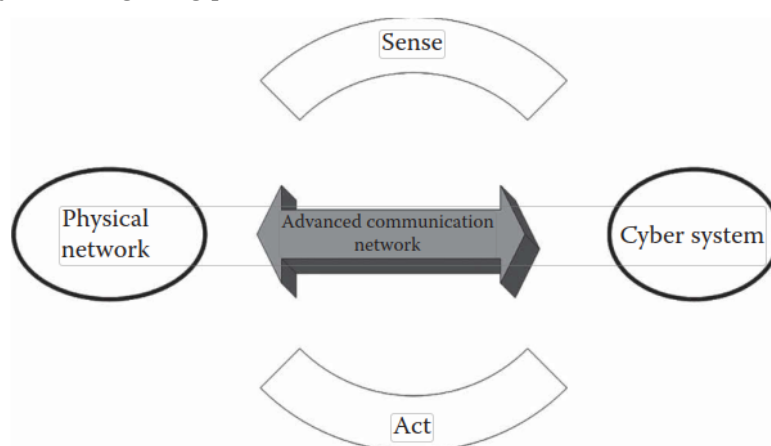


Fig: Smart grid as a cyber–physical system. Adapted from Yu and Xue (2016).

- **Key Communication Requirements:**
- **Latency** – network delay must be minimal.
- **Bandwidth** – critical for choosing wired/wireless tech.
- **Interoperability/Flexibility** – systems must work together.
- **Data Throughput** – max info transfer capacity.
- **Cybersecurity** – authentication, authorization, privacy.

- **Data Characteristics** – IoT devices generate big data: heterogeneous, varied, unstructured, noisy, redundant.

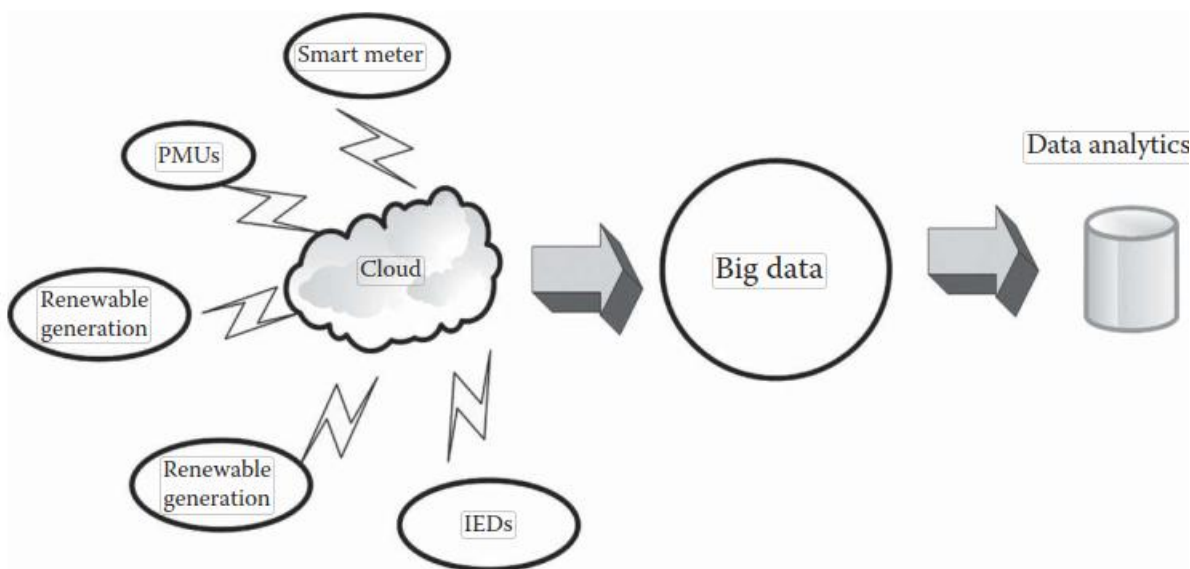


Fig : The integration of IoT and Big Data technology. Adapted from Marjani et al. (2017).

- **Technology Choices** – Must be reliable, secure, cost-effective, and adaptable to future needs.
- **Wired** – high data rate, costly, low mobility.
- **Wireless/Mobile** – scalable, flexible, lower cost, supports **M2M communication**.
- **Mobile Standards** – LTE-based enhancements for IoT.
- **NB-IoT (3GPP Release 13 & 14)** – efficient spectrum use, wide coverage, low-cost devices, high capacity, supports massive machine-to-machine (M2M) communication.

2.4 Data Traffic Pattern in a Smart Grid Environment

- Traditional power systems are **unidirectional**, with limited intelligence and no real-time communication (Bouhafs et al., 2012).
- Smart grids require **advanced communication systems** for protection, monitoring, and control.
- Existing systems like **SCADA** and **AMR** are inadequate for future smart grid demands (Lai and Lai, 2015).
- Future grids need **bidirectional communication**, ensuring robustness, reliability, and security.

Main functions (Marjani et al., 2017; Fan et al., 2012):

1. **Distribution Control & Protection** – IED devices detect/locate faults and exchange reporting messages.
2. **Wide Area Monitoring System** – Collects info from substations and large areas for critical decisions.
3. **Demand Response (DR)** – Manages multiple distributed energy resources with intermittent variables.

4. **Advanced Metering Infrastructure (AMI)** – Smart meters for billing, consumer interaction, load control, DR, and islanding detection.
 - Each application has **different latency and data requirements** (Kuzlu et al., 2014):
 - **Protection:** 1–10 ms, few bytes
 - **Control:** ~100 ms, few bytes
 - **Monitoring:** ~1 s, few–medium bytes
 - **Metering:** minutes–hours, medium size
 - AMI enables functions like **choosing energy sources, demand monitoring, and dynamic billing** (Sánchez-Ayala et al., 2013).
 - Smart grids will produce **huge amounts of data**, e.g., PMUs (sampled at 10–60 per cycle) and AMI (every 1–15 min).
 - **Big data analytics** is crucial for maintaining reliability, robustness, and grid modernization (Daki et al., 2017).

Phasor Measurement Units (PMUs)

- PMUs placed strategically perform **accurate voltage and current phasor measurements** using GPS (Liu et al., 2012).
- Data is sent to **Phasor Data Units (PDUs)** for processing.
- PMUs generate measurements per cycle (10, 20, 30, 60 most common).
- Required bandwidth depends on:
 - **BW = Nframe × fs × NPMU**
 - Nframe = frame size (bytes)
 - fs = sampling frequency
 - NPMU = number of PMUs connected
- **Higher sampling frequency = higher transmission capacity needs.**
- **Increasing PMU deployment = more synchrophasor data to analyze.**

Advanced Metering Infrastructure (AMI)

- Smart meters enable **numerous smart applications** when paired with reliable communication infrastructure.
- Each meter generates data with **different message size and sample rates** (Luan et al., 2010).
- Smart meters will be deployed in **residential, commercial, and industrial facilities**, creating massive data flow.
- **Estimating number of smart meters** depends on communication network design (coverage, antenna gain, propagation losses, geography).
- Formula for number of smart meters:

$$NSM = \rho \pi d^2$$

- ρ = smart meter density (per m²)
- d = coverage radius of base station (urban, suburban, rural variations).
- **Minimum data rate per smart meter = 64 kbps** for critical operations (Persia et al., 2015).

2.5 The Massive Flow of Information in a Smart Scenario

- **Massive data generation:**
 - Growth of **IEDs** (Intelligent Electronic Devices) → exponential data flow.
 - Utilities already generate **hundreds of millions of gigabytes**, rising to **terabytes annually**.
 - **Synchrophasors alone** produce **hundreds of terabytes per year** (Asad et al., 2017).
- **Opportunities from big data:**
 - Improves **grid reliability** and enables applications such as:
 - **Predictive analytics**
 - **Demand-side management**
 - **Real-time grid awareness**
 - **Outage detection**
 - **Asset management**
 - **Theft detection**

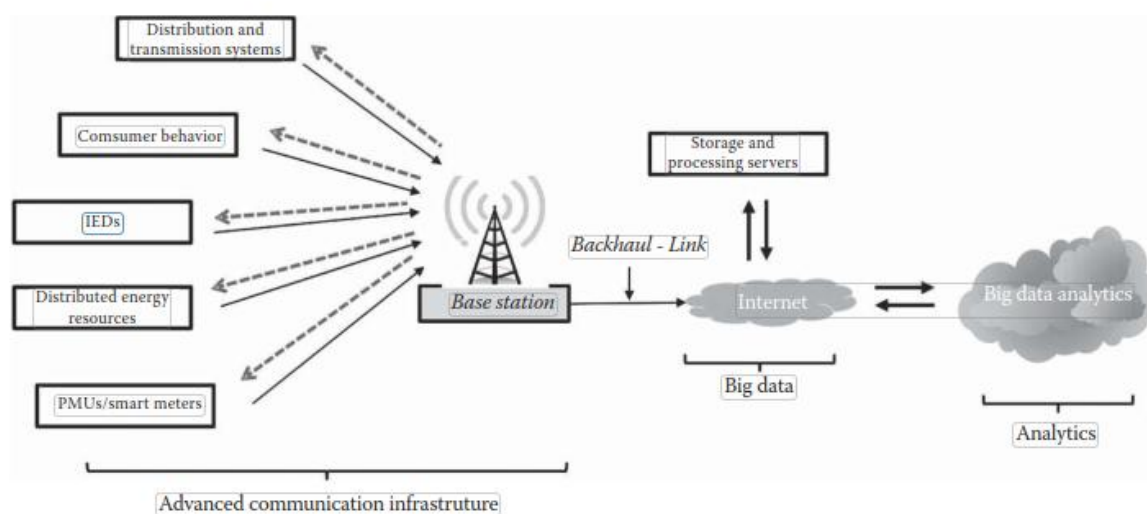


FIGURE 3.3

The integration of information and communication technologies in a smart grid context.

Role of big data (Jiang et al., 2016):

- Enhances **prediction, management, and processing** of grid operations.
- Recognizes **patterns in data** for better operational decisions.
- **Key application areas:**
 - **Demand Response (DR):** Predicting/analyzing user patterns → accurate demand forecasting.
 - **Distributed Energy Resources (DER):** Forecasting & scheduling intermittent energy sources for planning.
 - **AMI:** Smart meter data → customer behavior, load forecast, demand management.
 - **Distribution Automation:** Monitoring and sensing → predict outages, improve robustness.

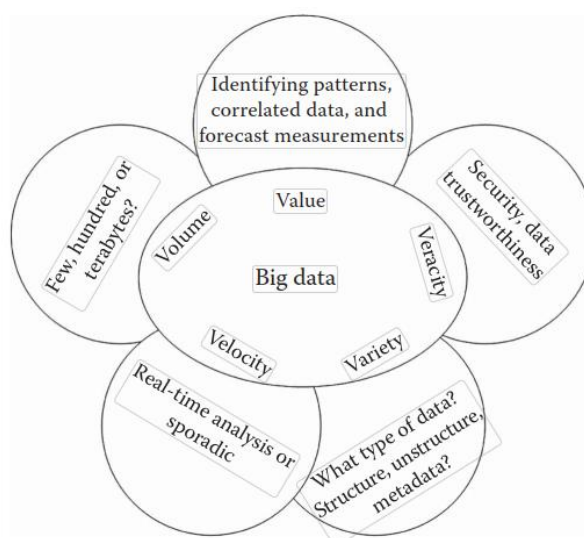


FIGURE 3.4
The 5 V's of big data in a smart grid scenario. Adapted from Subhani et al. (2015).

Big data foundation – 5 V's (Subhani et al., 2015):

- **Volume** – massive data scale
- **Velocity** – high-speed generation & transfer
- **Variety** – diverse data formats (structured/unstructured)
- **Value** – usefulness of data for decisions
- **Veracity** – data quality and accuracy

Advanced data analytics methods:

- Predictive analytics, data mining, AI, fuzzy theory

2.6 Volume of Generated Data in Smart Distribution – Case Study

General

- **Future smart grids = huge data** from IEDs, PMUs, and smart meters.
- In normal mode: periodic traffic (e.g., metering every 15 min).
- In critical mode: continuous messages, DR, load flow, outage info.

Case Study I – Data Generated by PMUs

Step-1: Setup

- IEEE 123 bus system is selected (high-load topology, widely used in research).
- PMU placement follows Jamil et al. (2014).
- **49 PMUs** deployed → optimal number for this system.

Step-2: Data Frame Structure

- Each PMU sends **frame packets** as per IEEE Std. C37.118 (2011).
- Packet = **80 fixed bytes** + fields (phasors, transducers, digital signals).
- Each PMU has **8 phasor channels** + **2 digital channels**.

Step-3: Sampling Rates

- Brazilian system frequency = **60 Hz**.
- Sampling rates considered: **10, 20, 30, 60 synchrophasors/sec**.

- Higher sampling rate = higher data generation.

Step-4: Bandwidth Requirement Formula

$$BW = N_{frame} \times f_s \times N_{PMU}$$

- Nframe = frame size (bytes)
- fs = sampling frequency
- NPMU = number of PMUs (49 in this case)

Step-5: Results

- At 10 samples/sec → data volume already in **hundreds of Mbps**.
- At higher rates → data grows drastically.

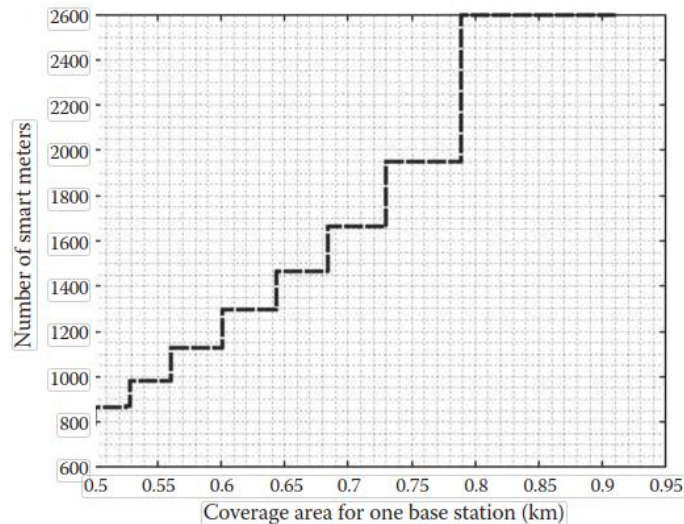
Step-6: Inference

- PMUs generate **massive continuous data streams**.
- Without **big data analytics**, it is difficult to handle, store, and process.
- Organized analysis = better grid performance and reliability.

Simulation: **IEEE 123 bus system** with NB-IoT communication.

Volume of Generated Data by PMUs

#PMU	#Bytes	Sampling Rate (Mbps)					
		10	12	15	20	30	60
1	112	8.96	10.752	13.44	17.92	26.88	53.76
2	224	16.06	19.27	24.08	32.11	48.17	96.34
3	336	43.16	51.79	64.74	86.32	129.48	258.96
6	672	232.02	278.43	348.04	464.05	696.07	1392.15
9	1008	1871.05	2245.26	2806.57	3742.10	5613.15	11226.29
10	1120	16764.60	20117.51	25146.89	33529.19	50293.79	100587.57
15	1680	225316.16	270379.39	337974.24	450632.32	675948.47	1351896.95
20	2240	4037665.548	4845198.66	6056498.323	8075331.097	12112996.6	24225993.29
30	3360	108532449.9	130238940	162798674.9	217064899.9	325597350	651194699.7
49	5488	4765008682	5718010419	7147513023	9530017364	1,4295E+10	28590052093
50	5600	2,13472E+11	2,5617E+11	3,20209E+11	4,26945E+11	6,4042E+11	1,28083E+12
60	6720	1,14763E+13	1,3772E+13	1,72144E+13	2,29526E+13	3,4429E+13	6,88577E+13



Case Study II – Data Generated by AMI (Smart Meters)

Step-1: Setup

- Smart meters deployed across the same **IEEE 123 bus system area** ($\approx 2 \text{ km}^2$, **urban scenario**).
- Meters randomly placed, exact optimal allocation not studied here.

Step-2: Data Rate Assumption

- Each smart meter requires **64 kbps** in critical mission scenarios (Persia et al., 2015).

Step-3: Coverage Analysis

- Communication via **NB-IoT wireless**.
- Factors: distance, propagation loss, antenna gain, coverage area.
- One base station coverage = **0.5 km^2** $\rightarrow$ supports **~ 800 meters**.

Step-4: Scaling to Whole System

- Total area = $\sim 2 \text{ km}^2$.
- Need **5 base stations** to cover entire system.
- Supports **~ 4000 smart meters**.

Step-5: Inference

- More smart meters = **exponential data growth**.
- Data must be processed by **advanced analytics** to avoid overload.
- Correct treatment ensures **efficient grid + communication flow**.

Inference

Case I (PMUs): Generate **huge continuous data**; needs big data analytics for reliability.

Case II (AMI): Thousands of smart meters $\rightarrow$ **exponential data**; requires **efficient communication & analytics**.

- **Overall:** Smart grids produce massive data; without proper treatment it's a burden, with analytics it ensures reliability and efficiency.

Outcomes

At the end of the module, students will be able to:

CO-1: Interpret the role of big data and machine-learning methods applicable to power systems and in particular to Smart Grid communications. [L2]

TEXT BOOKS:

Big Data Analytics in Future Power Systems, Ahmed F. Zobaa and Trevor J. Bihl, CRC Press 2019. 2019.

Reference Books/ Link

1. **Big Data Analytics for Power Systems** – [Big Data Analytics in Power Systems](#)
2. **Application of Big-Data Analytics in Power System Protection**-[Lec-37: Application of Big-Data Analytics in Power System Protection](#)