Vol. 03, Iss. 3 (2026) 191-209, DOI: 10.61552/JSI.2026.03.006

Journal of Sustainable Development Innovations

www.jsi.aspur.rs

Research article

Big Data Analytics Framework for Renewable Energy Integration and Smart Grid Decision Support in Ghana

Effraim Patrick[a,*], Budu Bismark[a], Mensah Michael[a]

[a]*Department of Electrical and Computer Engineering, Academic City University, Accra, Ghana.*

Keywords:

Big Data, Renewable energy integration, Smart grid, Machine learning, Solar forecasting, Ghana, Decision support, Apache Spark, LSTM, Random Forest.

* Corresponding author:

Effraim Patrick
E-mail: Effraimpat@gmail.com

Received: 22 June 2026
Revised: 26 July 2026
Accepted: 28 August

A B S T R A C T

The growing adoption of renewable energy sources, especially photovoltaic (PV) solar energy systems, introduces numerous difficulties for power systems due to generation intermittency, rising demands for electricity consumption, and the necessity for intelligent decision-making. In this regard, developing nations such as Ghana experience particular difficulties in deploying renewable energy sources that require sophisticated data management and analysis processes. Therefore, this article offers a conceptual Big Data analytics framework for the renewable energy sources integration and smart-grid decision-making in the context of the Ghana electricity industry. The suggested framework involves seven layers, which include data acquisition, data ingestion, distributed storage, distributed computing, analytics, visualization, and governance, among others, relying on such technologies as Apache Kafka, Hadoop Distributed File System (HDFS), Apache Cassandra, and Apache Spark. The proposed framework suggests that the machine learning algorithms such as Random Forest, Long Short-Term Memory (LSTM), and Isolation Forest can be applied to the future prediction of solar photovoltaic energy, future prediction of electricity demand and anomaly detection, respectively. A preliminary illustration using publicly available Ghanaian electricity demand data demonstrates the practical application context of the proposed framework; however, no machine learning models are implemented, trained, or experimentally validated in this study. The proposed framework provides a reference architecture to guide future implementation, empirical validation, and the development of data-driven smart-grid solutions for Ghana and other developing-country electricity systems.

1. INTRODUCTION

The global energy transition, driven by the Paris Agreement and other international climate initiatives, has accelerated the adoption of renewable energy technologies, particularly solar photovoltaic (PV) and wind power, as key components of sustainable electricity generation. As more than 140 countries pursue net-zero carbon emission targets by 2050, integrating variable renewable energy (VRE) sources into existing power systems has become a major technical and operational challenge [1]. Although renewable energy offers significant environmental and economic benefits, its intermittent nature introduces uncertainties in power generation, making grid stability, demand balancing, and operational planning increasingly complex.

In Ghana, electricity generation has traditionally relied on hydroelectric and thermal power plants. However, rapid urbanization, population growth, and industrial expansion have significantly increased electricity demand, placing considerable pressure on the national power system [2]. To diversify its energy mix and improve energy security, Ghana has invested in renewable energy projects, including the 55 MW Bui Solar Power Plant and the 28 MW Kaleo Solar Project. The country's installed solar capacity has reached approximately 130 MW, with further expansion planned under the Renewable Energy Master Plan, which targets a 10% renewable energy contribution by 2030 [2, 3]. Despite these investments, integrating solar PV into the national grid remains challenging because electricity generation is highly dependent on solar irradiance, weather variability, seasonal Harmattan conditions, and fluctuations in electricity demand.

The increasing deployment of smart meters, Supervisory Control and Data Acquisition (SCADA) systems, Internet of Things (IoT) sensors, and weather-monitoring stations has transformed modern electricity networks into data-intensive environments. These technologies continuously generate large volumes of heterogeneous and high-velocity operational data that exceed the capabilities of conventional data-processing approaches. Consequently, Big Data technologies have emerged as an effective solution for managing, processing, and analyzing smart-grid data to support real-time monitoring, renewable energy forecasting, and operational decision-making. Distributed computing platforms such as Apache Kafka, Apache Spark, Hadoop Distributed File System (HDFS), and Apache Cassandra have demonstrated significant potential for scalable smart-grid analytics and renewable energy applications in several countries, including Kenya, Nigeria, South Africa, Morocco, and Egypt [4, 5].

Recent research has investigated Big Data platforms, renewable energy forecasting techniques, smart-grid architectures, and machine learning methods for intelligent energy management. However, these studies have generally focused on individual aspects such as forecasting algorithms, distributed computing platforms, or smart-grid applications in isolation. There remains a lack of integrated conceptual frameworks that combine Big Data infrastructure, distributed computing technologies, machine learning components, data governance, and decision-support mechanisms within a unified architecture tailored to the needs of developing countries such as Ghana. Furthermore, limited research has considered the practical challenges associated with implementing such frameworks in the Ghanaian electricity sector, including infrastructure readiness, data management, cybersecurity, and regulatory considerations.

To address this gap, this paper proposes a conceptual Big Data analytics framework for renewable energy integration and smart-grid decision support in Ghana.

The proposed framework is intended as a reference architecture for future implementation rather than an operational or experimentally validated system.

It integrates distributed computing technologies, conceptual machine learning modules, and data management processes into a unified architecture capable of supporting scalable, data-driven decision-making for renewable energy integration and smart-grid operation.

The proposed framework makes the following contributions:

- Proposes a seven-layer conceptual Big Data architecture that integrates data acquisition, ingestion, distributed storage, distributed processing, analytics, visualization, and governance.
- Combines distributed stream-processing technologies (Apache Kafka, Apache Spark, HDFS, and Apache Cassandra) with conceptual machine learning modules within a unified smart-grid framework.
- Introduces a Ghana-specific framework that explicitly considers infrastructure limitations, smart-meter deployment challenges, data quality issues, cybersecurity requirements, and regulatory constraints relevant to developing-country electricity systems.
- Identifies Random Forest, LSTM, and Isolation Forest as suitable candidate models for future applications in solar PV forecasting, electricity demand forecasting, and grid anomaly detection.
- Provides a phased implementation roadmap that links conceptual design to future prototype development and empirical validation.
- Demonstrates the practical application context of the framework through an illustrative analysis of publicly available Ghanaian electricity demand data, while clearly distinguishing this illustration from model implementation or validation.

The remainder of this paper is organized as follows. Section 2 reviews the literature on renewable energy integration, smart-grid technologies, Big Data analytics, and machine learning applications. Section 3 presents the research methodology and problem definition. Section 4 describes the proposed conceptual framework and its supporting data architecture. Section 5 discusses the conceptual machine learning components and analytical workflow. Section 6 presents the Ghanaian implementation context and preliminary data illustration.

Section 7 discusses implementation challenges, opportunities, and section 8 policy recommendations. Finally, Section 9 concludes the paper by summarizing the study and outlining directions for future implementation and empirical validation.

2. LITERATURE REVIEW

2.1 Renewable Energy Integration Challenges in Sub-Saharan Africa

The integration of renewable energy into power generation systems in sub-Saharan Africa presents distinct challenges compared to highly digitized power systems in developed countries. In many developing countries, key constraints include inadequate grid infrastructure, limited forecasting capability, insufficient grid flexibility, and weak real-time monitoring and control systems. According to the International Renewable Energy Agency (IRENA), increased penetration of variable renewable energy (VRE) sources such as solar and wind requires significant improvements in grid flexibility, forecasting accuracy, and real-time operational visibility to ensure system reliability and stability [6].

Recent studies in African energy systems highlight the growing relevance of data-driven approaches in addressing these challenges. In Kenya, research has demonstrated that artificial intelligence-based forecasting techniques can improve renewable energy prediction accuracy, thereby reducing uncertainty associated with solar and wind intermittency [5]. Similarly, studies on smart grid systems in developing countries emphasize the importance of distributed data processing architectures, sensor networks, and advanced analytics tools for managing large-scale energy datasets [7].

However, despite these advancements, most existing studies remain limited to pilot-scale implementations or simulation based environments, with minimal deployment in operational grid systems across sub-Saharan Africa. This highlights the need for scalable, data-centric frameworks tailored to the infrastructural and operational realities of emerging power systems in the region.

2.2 Data Applications in Smart Grid Management

In recent years, big data technologies have gained significant attention in smart grid management and optimization. Modern smart grid infrastructures generate large volumes of heterogeneous data from multiple sources, including smart meters, Supervisory Control and Data Acquisition (SCADA) systems, Internet of Things (IoT) devices, and renewable energy forecasting systems. These data streams are characterized by high volume, velocity, and variety, necessitating scalable data storage and processing architectures.

To address these requirements, distributed computing frameworks such as Apache Hadoop, Apache Spark, and Apache Kafka have been widely adopted for real-time data ingestion, processing, and analytics in smart grid applications. These frameworks enable parallel processing of large datasets and support near real-time decision-making in energy systems.

Furthermore, several studies have demonstrated that machine learning techniques can significantly improve forecasting accuracy and grid management performance. Algorithms such as Random Forest, Gradient Boosting Machines, and Long Short-Term Memory (LSTM) networks have been widely applied in electricity demand forecasting and renewable energy prediction tasks [8, 9]. These approaches generally outperform traditional statistical methods due to their ability to model nonlinear relationships and temporal dependencies in energy data.

However, much of the existing literature is based on datasets and infrastructure from highly developed power systems, where data quality and availability are relatively high. Consequently, the direct applicability of these approaches to sub-Saharan African power systems remains limited due to differences in infrastructure maturity, data reliability, and operational constraints.

Recent studies in African contexts have begun exploring artificial intelligence applications in renewable energy forecasting and smart grid optimization. For example, Morara and Ondieki [5] emphasize the relevance of AI-based approaches for improving energy forecasting and supporting sustainable energy planning in Kenya. Additionally, research on distributed computing architectures indicates that scalable big data systems can enhance the management of large-scale energy operational data in developing-country contexts .

Despite these developments, there remains a clear gap in the literature regarding fully integrated big data frameworks that combine data ingestion, pre-processing, forecasting, anomaly detection, and decision-support capabilities specifically tailored to African power systems such as Ghana.

2.3 Machine Learning in Energy Forecasting

Machine learning techniques have gained increasing attention in energy forecasting due to their ability to model complex nonlinear relationships and temporal dependencies in large operational datasets. Commonly used approaches include ensemble learning methods, deep learning architectures, and unsupervised anomaly detection techniques.

Random Forest, an ensemble learning method based on bootstrap aggregation and random feature selection, has demonstrated strong performance in handling high-dimensional and noisy energy datasets [8]. Its robustness to overfitting and ability to model nonlinear relationships make it suitable for energy forecasting applications.

Deep learning models, particularly Long Short-Term Memory (LSTM) networks, are widely used for sequential forecasting tasks due to their ability to capture long-term temporal dependencies in time-series data [9]. In smart grid applications, LSTM models have been applied to electricity demand forecasting, renewable energy prediction, and the analysis of seasonal and short-term consumption patterns [10].

In addition, anomaly detection methods have become increasingly important in smart grid monitoring and cybersecurity. Isolation Forest, introduced by Liu et al. [11], provides an efficient unsupervised learning approach for detecting anomalies such as sensor failures, abnormal voltage fluctuations, and irregular energy usage patterns without requiring large labeled datasets.

2.4 Research Gap

Although significant research has been conducted on big data analytics for smart grid management and renewable energy forecasting, most existing studies focus on individual components of the energy analytics pipeline, such as forecasting algorithms, distributed computing platforms, or smart grid applications, rather than providing integrated end-to-end frameworks. Consequently, limited attention has been given to comprehensive architectures that combine data engineering, distributed computing, machine learning, data governance, cybersecurity, and decision-support systems within a unified framework for power system management [5, 12].

In the context of Ghana, existing energy research has primarily concentrated on renewable energy deployment, infrastructure expansion, and generation capacity planning. Comparatively fewer studies have explored how Big Data technologies can be integrated with predictive analytics and intelligent decision-support mechanisms to support renewable energy integration and smart grid operations under the infrastructural and operational constraints typical of developing-country power systems [3].

Therefore, there remains a need for scalable, context-aware, and implementation-oriented conceptual frameworks that integrate heterogeneous data sources, distributed data processing, machine learning, and operational decision support into a single architecture tailored to emerging electricity systems.

To address this gap, this study proposes a novel conceptual Big Data analytics framework for renewable energy integration and smart grid decision support in Ghana. Unlike existing frameworks that typically emphasize isolated analytical techniques or computing platforms, the proposed framework integrates seven functional layers comprising data acquisition, data ingestion, distributed storage, distributed processing, machine learning analytics, visualization, and governance within a unified reference architecture. In addition, the framework explicitly incorporates Ghana-specific deployment considerations, including infrastructure limitations, data quality challenges, cybersecurity requirements, regulatory considerations, and a phased implementation roadmap. Rather than presenting an implemented analytical system, the framework serves as a reference architecture to guide future prototype development, empirical validation, and practical deployment in Ghana and other developing-country electricity systems.

3. PROBLEM DEFINITION AND MOTIVATION

The incorporation of renewable-energy resources into the Ghanaian national electric power network poses several technical difficulties. The solar photovoltaic (PV) technology, which happens to be one of the most rapidly developing renewable energy resources in Ghana, is an inherently intermittent power source, as its production of electricity depends highly on climatic factors, including solar radiation, clouds, concentration of dust in the air during the Harmattan period, solar panel temperature, and humidity levels. Unlike hydro and thermal power plants, PV power is not dispatch able based on electric power demand, as it can vary within a short interval of time due to weather changes [4, 13].

On the other hand, energy demand in Ghana shows great variations both daily and seasonally. The highest electricity demand is normally recorded during evenings when there is little production by solar generators, thereby making it difficult to balance between supply and demand. Other circumstances that can increase electricity demand are during national holidays and times when industries use more electricity, hence putting strain on reserve electricity producers and the entire grid.

All the variables mentioned, together with the lack of effective real-time analysis tools and the huge amounts of operational data generated from sensors, present a challenge for decision support to energy sector operators. Currently, the Electricity Company of Ghana and the Ghana Grid Company need a system that will be able to handle: (i) collection of operational data from multiple sources in real time; (ii) prediction of renewable energy production and electricity demand; (iii) detection of possible anomalies or fault within the grid network; and (iv) support data-driven decisions made in complex smart grids.

3.1 Scope and Delimitations

This study proposes a conceptual big data architecture framework for the integration of renewable energy resources into smart grid systems within the Ghanaian electricity sector. The framework is developed as a theoretical architectural model rather than a fully implemented or experimentally validated system.

The scope of the study is limited to the conceptual design of a big data-enabled smart grid framework, focusing on key components such as data acquisition, distributed data processing, machine learning integration, anomaly detection, and decision-support mechanisms. The study emphasizes system architecture and functional relationships rather than software implementation or hardware deployment.

Geographically, the study focuses on Ghana's national electricity grid, with particular emphasis on operational conditions in selected service areas of the Electricity Company of Ghana (ECG), including the Greater Accra, Ashanti, and Western regions. These regions are selected due to their high electricity demand and increasing penetration of distributed renewable energy systems [14].

This research does not include real-time field deployment, hardware installation, or utility-scale operational testing. Instead, it focuses on conceptual modelling and system design considerations necessary to support future implementation of scalable smart grid infrastructures in developing-country contexts

4. DATA SOURCES AND BIG DATA CHARACTERIZATION

4.1 Data Source Inventory

The proposed framework integrates heterogeneous data sources required for renewable energy forecasting, grid monitoring, and decision-support operations. These data sources originate from utility infrastructure, meteorological services, renewable energy installations, and supervisory monitoring systems. Their integration enables real-time analytics and supports intelligent decision-making within a distributed big data architecture.

Table 1 summarizes the heterogeneous data sources incorporated into the proposed framework. The table highlights the diversity of operational, meteorological, and renewable energy datasets required to support forecasting, anomaly detection, and smart-grid decision support. It also illustrates the different temporal resolutions and data providers, emphasizing the need for distributed storage and processing technologies capable of handling heterogeneous high-volume data streams.

Table 1. Data Sources and Operational Characteristics.

Data Source	Data Type	Granularity	Est. Daily Volume	Provider
Solar PV Plant Output	Structured numerical time-series	1-minute intervals	High-frequency generation telemetry from grid-connected solar facilities	ECG / Plant SCADA Systems
Smart Meter Readings	Structured time-series	15-minute intervals	Large-scale customer consumption data from the advanced metering infrastructure	ECG AMI Systems
Weather and Meteorological Data	Structured and semi-structured	Hourly / sub-hourly	Environmental and atmospheric monitoring datasets	Ghana Meteorological Agency / NASA POWER API
SCADA Grid Monitoring Data	Structured operational streaming data	Sub-second intervals	Real-time grid operational measurements and control signals	GRIDCo SCADA Systems

The diversity, velocity, and volume of such data collections reflect the essential characteristics of the Big Data domain. Specifically, smart grids continuously produce streams of data that can be effectively leveraged through distributed storage solutions, stream processing techniques, and machine-learning-based analysis.

4.2 Big Data Characteristics – 5Vs Framework

All datasets mentioned within the proposed framework possess certain basic characteristics that are typical for Big Data environments. Those characteristics have been traditionally characterized by the five-dimensional (5Vs) Big Data model that includes volume, velocity, variety, veracity, and value.

1. Volume: The current smart grid systems produce large volumes of data using their SCADA system, smart meters, renewable energy sources, and weather data collection devices. With prolonged use, those datasets become multi-terabytes of historical data that need to be stored in a scalable way.
2. Velocity: Datasets produced in the energy sector vary in terms of their update rate depending on the source. Measurements made using the SCADA system occur several times per second, while data captured by smart meters is collected periodically. Meteorological data is refreshed at an hourly and sub hourly frequency.
3. Variety: The model includes several types of heterogeneous data inputs, including

structured number series data, semi-structured log data, and unstructured imagery data, which is commonly employed in energy production prediction use cases. The heterogeneity in data type implies that there should be a need for more elastic data engineering and analysis pipelines.

4. Veracity: Typical power grid data used in operations usually includes some form of missing data, communication problems, noise in sensor data, and inconsistency in the calibration of the systems. There will thus be a need to employ pre-processing techniques such as outlier removal and normalization in data analysis.
5. Value: The overarching aim of Big Data analysis in smart grid technology involves enhancing decision-making, forecasting, system reliability, and integration efficiency. Increased accuracy in forecasting and anomaly detection could assist in decreasing operational uncertainties, managing reserves better, and optimizing renewable energy integration within the electricity industry in Ghana.

As illustrated in Figure 1, the five fundamental characteristics of Big Data-Volume, Velocity, Variety, Veracity, and Value are closely interconnected within smart-grid environments. These characteristics collectively justify the adoption of distributed computing technologies and advanced analytics to manage large-scale operational data generated by renewable energy systems and electricity networks.

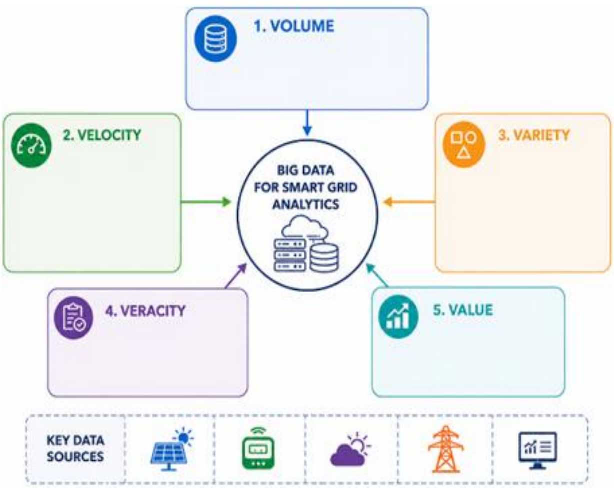

Fig. 1. The correlation between data volume, velocity, variety, veracity, and value.

4.3 Data Management Strategy for Future Implementation

4.3.1 Data Sources

The proposed framework is designed to integrate heterogeneous datasets originating from multiple operational and environmental sources within Ghana's electricity sector. These include smart meter measurements from the Electricity Company of Ghana (ECG), Supervisory Control and Data Acquisition (SCADA) data from the Ghana Grid Company (GRIDCo), meteorological observations from the Ghana Meteorological Agency (GMet), satellite-derived weather variables from the NASA POWER API, and solar photovoltaic operational data from grid-connected renewable energy installations. These data sources collectively provide the temporal, operational, and environmental information required for renewable energy forecasting, electricity demand prediction, anomaly detection, and decision support.

4.3.2 Dataset Availability

This study does not construct or analyse a dedicated machine learning dataset. Instead, the proposed framework identifies publicly available and utility-generated data sources that could support future implementation. Public datasets, including historical electricity demand records published by GRIDCo and meteorological observations from the NASA POWER platform, are used solely to illustrate the applicability of the proposed framework.
Future implementation would require access to utility operational databases, smart meter measurements, SCADA records, and renewable energy generation data collected over extended periods.

4.3.3 Conceptual Data Pre-processing Workflow

Within the proposed framework, raw operational data are assumed to undergo standard pre-processing before analytics and machine learning are performed. Typical pre-processing operations include:

- synchronization of heterogeneous time-series collected at different sampling intervals;

- removal of duplicate and corrupted records;
- treatment of missing observations using appropriate interpolation or imputation techniques;
- outlier detection and filtering;
- normalization or standardization of numerical variables;
- feature engineering, including temporal variables, lag features, rolling statistics, and weather-derived indicators.

These pre-processing operations are included conceptually to ensure that future machine learning models receive consistent, high-quality input data.

4.3.4 Conceptual Validation Strategy

No machine learning models are trained or validated in this study. Nevertheless, the proposed framework anticipates that future implementations will adopt chronological data partitioning to preserve temporal dependencies in time-series forecasting applications. Model performance may then be evaluated using standard forecasting metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). For anomaly detection models, evaluation may additionally employ Precision, Recall, F1-score, Receiver Operating Characteristic (ROC) curves, and Area Under the Curve (AUC), depending on the availability of labelled operational events.

4.3.5 Expected Deployment Requirements

Future implementation of the proposed framework would require a distributed computing environment capable of supporting continuous data ingestion, storage, analytics, and visualization. The conceptual architecture assumes the use of Apache Kafka for streaming data ingestion, Hadoop Distributed File System (HDFS) for distributed storage, Apache Cassandra for high-frequency operational data, Apache Spark for distributed analytics, and visualization platforms such as Apache Superset or Grafana. Deployment would further require reliable network connectivity between field devices and utility control centres, sufficient computing resources, secure communication channels, role-based access control, and compliance with Ghana's Data Protection Act and applicable cybersecurity policies.

5. PROPOSED FRAMEWORK ARCHITECTURE

5.1 Overview of Architecture

The suggested model relies on a hierarchical architecture based on the concept of Lambda Architecture, developed by Marz and Warren [15]. As the name implies, the architecture allows combining batch and stream processing to deliver a solution capable of delivering scalable data analysis solutions. In order to meet the particular needs of smart grids, another layer dedicated to governance and security is included in the hierarchy for providing proper data protection, access, compliance, and cybersecurity.

Overall, the architecture includes seven interrelated layers that can be used to forecast renewable energy, predict electricity demand, detect anomalies, and provide support for making decisions based on Big Data and machine learning. The seven layers of the proposed architecture are as follows:

- Data Collection
- Data Ingestion and Streaming
- Storage
- Processing and Transformation
- Analytics and Machine Learning
- Visualization and Decision Support

Each layer has an assigned operational responsibility in the entire data pipeline process and interoperates with other neighbouring parts. The framework allows for the gathering of different smart grid datasets, data streaming, distributed storage, big data processing, prediction, and visualization decision support under one architecture.

Figure 2 presents the overall conceptual architecture of the proposed framework. It illustrates the end-to-end flow of information from heterogeneous data acquisition through data ingestion, distributed storage, processing, machine learning analytics, visualization, and governance. The figure demonstrates how these functional layers interact to support scalable renewable energy forecasting and smart-grid decision support within the Ghanaian electricity sector.

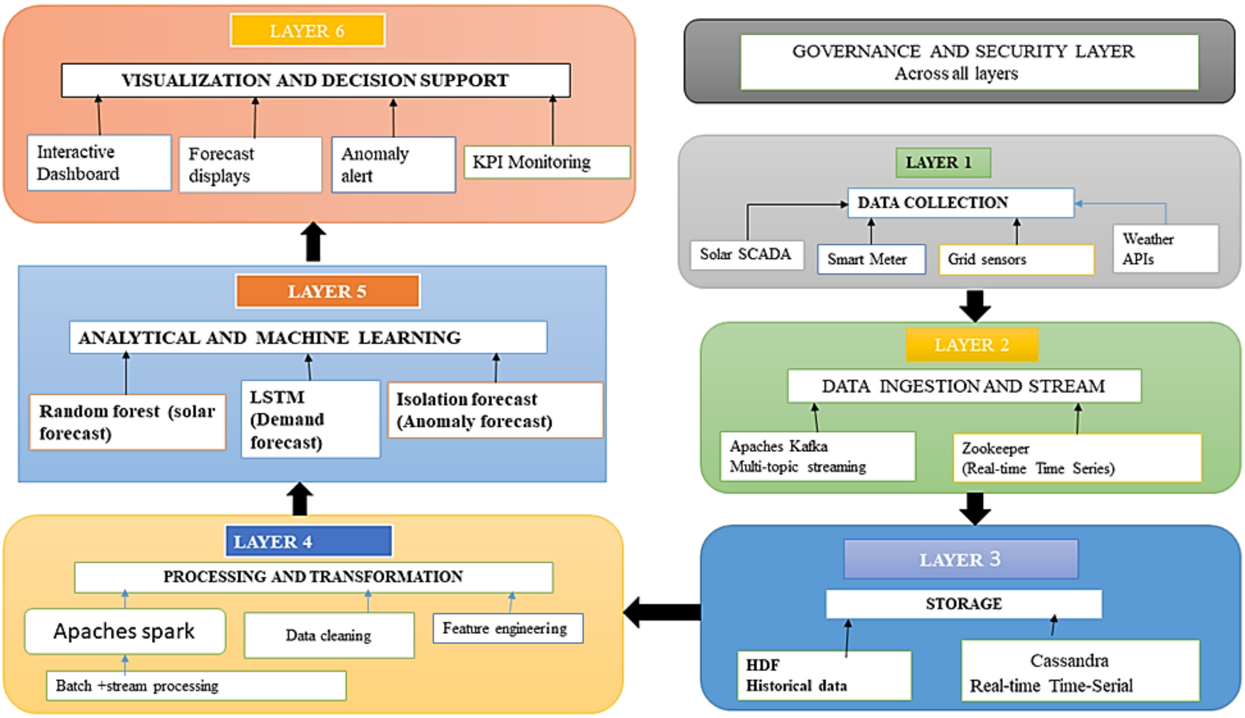

Fig. 2. Proposed Six-Layer Big Data Framework for Smart Grid Analytics and Renewable Energy Generation in Ghana.

5.2 LAYER-BY-LAYER SPECIFICATION

5.2.1 Data Collection Layer

This is the layer that consists of both physical and virtual infrastructures capable of providing raw operational data from the proposed smart-grid system. High-frequency data such as generation, temperature, and efficiency from solar photovoltaic monitoring systems are collected using industry communication standards, including Modbus TCP and IEC 61850**.** Smart meters under the AMI collect energy-consumption data periodically using different forms of communication like GPRS and NB-IoT. Moreover, the transmission grid SCADA managed by GRIDCo provides real-time operational data such as voltage and frequency.

The meteorological data can be sourced from the Ghana Meteorological Agency (GMetS). Moreover, the meteorological data can also be augmented by using satellite-based environment data sources from the NASA POWER API. The various types of weather data include information such as solar radiation levels, temperatures, humilities, wind speeds, and cloud cover, among others.

5.2.2 Data Ingestion Layer – Apache Kafka

Apache Kafka is proposed as the primary distributed streaming platform within the ingestion layer in the context of the data ingestion layer. This framework segregates the received operational data into various streaming topics, which include solar power generation, smart meter demand, SCADA operational, and meteorological data streams. The decoupling of data producers from consumers is made possible by virtue of the publishing-subscription model adopted by Apache Kafka, leading to improved system scalability and fault tolerance.

The data ingestion layer caters to real-time high-throughput streaming processes, specifically for smart grids, and ensures that the system remains resilient even in the presence of network and node disruptions.

5.2.3 Storage Layer – HDFS & Apache Cassandra

The storage layer leverages a hybrid distributed storage architecture that integrates Hadoop Distributed File System (HDFS) for historical batch storage along with Apache Cassandra for real-time time-series data processing. HDFS is used for archival storage of processed operational datasets and machine learning training data owing to its scalable and distributed fault tolerance capabilities. Processed datasets will be stored using analytical formats such as Apache Parquet to enable effective compression and efficient querying. Apache Cassandra will be used for real-time operations that involve high-frequency grid monitoring data and stream processing. The database architecture is geared towards time-series applications and enables scalable writes that are typical in smart grids.

5.2.4 Processing Layer – Apache Spark

The Apache Spark framework includes features for both large-scale batch processing and near-real-time stream processing. The batch processing functionalities are utilized for tasks such as historical data analysis, feature generation, training of machine learning models, and regular performance reports. In addition, Spark Structured Streaming is conceptually integrated to support near-real-time processing from the smart grids in order to make short-term forecasts and detect anomalies.

The architecture of the processing layer was designed in such a way that it can run in a distributed computing environment. This would allow for large-scale data analysis within the scope of the operations of Ghana's power utility company. The proposed solution can be scaled gradually using available computing resources.

5.2.5 Visualization and Decision Support Layer

The visualization and decision-support layer is designed to provide an interactive interface for grid managers and energy sector stakeholders through dashboard tools such as Apache Superset and Grafana. Within the proposed framework, these tools are intended to support the visualization of key operational metrics derived from smart grid data streams [16, 17].

These dashboards are conceptually designed to present near real-time and forecasted information on renewable energy generation, electricity demand, supply–demand balance, grid stability indicators, and anomaly detection alerts. This enables improved situational awareness for grid operation and planning.

The interface is conceptualized for potential use by operators within Ghana's electricity sector and is aligned with typical enterprise-level energy management systems. Furthermore, the design incorporates role-based access control mechanisms to support secure and coordinated data access among relevant stakeholders, including the Electricity Company of Ghana (ECG), the Ghana Grid Company (GRIDCo), and the Energy Commission.

6. MACHINE LEARNING MODELS: SPECIFICATION AND CONCEPTUAL ANALYSIS

6.1 Model 1: Solar PV Output Prediction using Random Forest Regressor

6.1.1 Rationale for Model Choice and Model Structure

Accurate predictions for solar PV energy require intelligent machine learning algorithms that can capture the nonlinear dynamics between the environmental variables and power generation effectively. The environmental factors that affect solar energy production in Ghana include, but are not limited to, solar irradiation, ambient temperature, relative humidity, cloudiness, and the amount of dust in the atmosphere during the Harmattan period [18].

Random Forest (RF) algorithm was chosen as the main model for the forecasting purpose because of its efficiency in dealing with noisy and large-dimensional data sets, lack of over-fitting problem, and better predictions [8].

In the context of the suggested intelligent grid system, it is envisaged that the model will be used to make solar power forecasting in the short term with the help of operational data, meteorological information, and time variables. The solar PV production forecast is calculated through the aggregation of predictions made from all the decision trees forming the forest. Formally:

$$p_{PV} = \frac{1}{N}\sum_{i=1}^{N} T\, i\, (x) \tag{1}$$

where p_{PV} is the estimated photovoltaic power generation, (N) is the number of total decision trees in the ensemble, $T_{i(X)}$ is the prediction produced by the $i^{\{th\}}$ decision tree, and (X) refers to the input feature vector containing weather-related variables along with operational and temporal features. Through the averaging effect of predictions from several trees, the model can minimize the variance and better represent the nonlinear relationship between the weather conditions and power production [19].

The input feature vector used for solar power forecasting can be represented as:

$$X = \left[GHI, DNI, T_{\{amb\}}, T_{\{panel\}}, RH, AOD, h, d, P_{\{t-1\}}, P_{\{t-3\}}, GP_{\{t-3\}}, CCI\right] \quad (2)$$

where (GHI) denotes Global Horizontal Irradiance, (DNI) is Direct Normal Irradiance, $T_{\{amb\}}$ represents ambient temperature, $T_{\{amb\}}$is photovoltaic panel temperature, (RH) denotes relative humidity, (AOD) represents Aerosol Optical Depth, (h) denotes the hour of the day, (d) represents the day index, $P_{\{t-1\}}$ and $P_{\{t-3\}}$ are lagged solar power generation values, (G$P_{\{t-3\}}$ represents rolling-average solar irradiance, and (CCI) denotes the Cloud Cover Index

The selected input variables were chosen based on their established influence on photovoltaic power generation and their availability within modern smart-grid environments.
Solar irradiance variables (GHI) and (DNI) directly affect energy production, while ambient and panel temperatures influence photovoltaic conversion efficiency. Relative humidity, cloud cover, and aerosol concentration are particularly important in the Ghanaian context due to seasonal atmospheric variations, including Harmattan dust events.

Historical generation measurements,$P_{\{t-1\}}$,$P_{\{t-3\}}$ and rolling irradiance indicators GHI_{roll} provide temporal information that enables the model to capture short-term generation patterns and improve forecasting accuracy.

6.1.2 Feature Engineering

The input feature vector for the proposed Random Forest model incorporates meteorological variables, operational measurements, temporal indicators, and historical lag features associated with photovoltaic power generation behaviour.

Table 2 summarizes the meteorological, operational, and temporal variables proposed as inputs to the Random Forest model. The selected features reflect factors known to influence photovoltaic power generation and provide the conceptual basis for future model development and empirical validation

Table 2. Random Forest Feature Vector for Solar Power Forecasting.

Feature	Symbol	Unit	Source	Engineering Description
Global Horizontal Irradiance	GHI	W/m^2	GMetS / NASA POWER	Direct solar irradiance measurement
Direct Normal Irradiance	DNI	W/m^2	NASA POWER API	Satellite-derived irradiance estimate
Ambient Temperature	$T_a$mb	°C	GMetS Station	Near-surface atmospheric temperature
Panel Surface Temperature	$T_p$anel	°C	PV SCADA Sensors	Solar module operating temperature
Relative Humidity	RH	%	GMetS Station	Atmospheric humidity level
Aerosol Optical Depth	AOD	Dimensionl ess	MODIS Satellite Data	Harmattan dust concentration proxy
Hour of Day	h	Integer	Derived	Cyclic temporal encoding
Day of Year	d	Integer	Derived	Seasonal cyclic encoding
Previous Output (t−1)	P_{t-1}	MW	SCADA System	One-step lag generation feature
Previous Output (t−3)	P_{t-3}	MW	SCADA System	Multi-step lag generation feature
Rolling Mean Irradiance	GHI_roll	W/m^2	Derived	Moving-average irradiance feature
Cloud Cover Index	CCI	%	Satellite / GMetS	Estimated atmospheric cloud coverage

6.1.4 Training Configuration and Dataset Partitioning

The proposed Random Forest model is conceptually designed to utilize a dataset comprising solar photovoltaic generation data and meteorological variables recorded at hourly intervals over an extended time period. Within the framework, data preprocessing and feature preparation are assumed to follow standard practices commonly adopted in time-series forecasting systems.

To preserve temporal integrity in energy forecasting applications, the conceptual framework assumes that chronological ordering of data is maintained, and that data leakage between future and past observations is strictly avoided. This is essential in renewable energy forecasting scenarios where temporal dependencies significantly influence model behaviour.

This design is intended to support future implementation and empirical validation within operational smart grid environments.

6.1.5 Expected Performance Characteristics

Within the proposed framework, the Random Forest model is conceptually intended to support short-term solar photovoltaic power forecasting using meteorological and operational input variables. The ensemble-based nature of the model enables it to represent nonlinear relationships between environmental conditions and solar power generation.

In related literature, ensemble learning approaches have been widely reported as effective for solar forecasting due to their ability to capture complex interactions among weather-related variables. Accordingly, the Random Forest model is adopted in this study as a suitable candidate for inclusion within the proposed big data-driven smart grid architecture.

The actual performance evaluation and validation of the model, including metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and coefficient of determination (R^2), are reserved for future empirical implementation studies.

6.2 Model 2: Electricity Demand Forecasting - Lstm Neural Network

6.2.1 Model Rationale

Predicting the future demand for electricity plays an integral role in a smart grid as it helps to plan electricity generation, reserves, incorporate renewable energy sources, and make decisions. Demand is affected by time-based, weather-related, and economic factors that contribute to the complexity of the demand sequence changing over time.

The Long Short-Term Memory (LSTM) networks are used in this context due to their ability to account for temporal patterns in sequences. Several studies have already established the potential of using LSTM networks for short-term load forecasting due to their capacity to store information from the past. [9, 10].

6.2.2 Input Features

Several variables must be considered when forecasting electricity demand, as they affect patterns over time. The proposed LSTM architecture combines these inputs to predict the dynamic nature of energy usage behaviour. The proposed set of input features is given by:

$$X_t = D_{\{t-1\}}, D_{\{t-2\}}, D_{\{t-24\}}, Temp, RH, Hour, DOW, Month, Holiday \quad (3).$$

where $D_{\{t-1\}}, D_{\{t-2\}}$and $D_{\{t-24\}}$ are lagged observations of the electricity demand variable, (Temp) is the ambient temperature, (RH) is the relative humidity level, (Hour) is the time of the day, (DOW) is the day of the week, (Month) is season-related information, and (Holiday) is an indicator of any special events on a given day.

The use of past demand variable values helps the model identify patterns of repetitive behaviour of electricity consumption. The other explanatory variables allow for considering additional context related to consumer behaviour and seasonality. The use of such input features has become a common practice in LSTM-based load forecasting research since these features enable learning temporal dynamics along with other affecting factors [10, 16].

6.2.3 Proposed Network Architecture

The proposed model for the prediction of electricity demand would involve the application of the Stacked LSTM network architecture that is capable of capturing the relationships and non-linearity's present in the data concerning the consumption of electricity. The proposed network architecture involves the use of two LSTM layers, followed by a dense fully-connected layer and the final output layer that performs the function of forecasting the electricity demand.

In order to enhance the generalizability of the network architecture and avoid the phenomenon of overfitting in the future applications of the proposed network architecture, the process of dropout has been utilized between LSTM layer [9].

Table 3 outlines the conceptual architecture of the proposed LSTM network for electricity demand forecasting. The layered configuration is intended to capture temporal dependencies in electricity consumption while maintaining a model structure suitable for future implementation within distributed smart-grid environments.

Table 3. Proposed LSTM Network Architecture for Demand Forecasting.

Layer	Type	Configuration	Activation
Input Layer	Sequential Input	Historical multivariate time-series window	—
LSTM Layer 1	LSTM + Dropout	128 hidden units	tanh
LSTM Layer 2	LSTM + Dropout	64 hidden units	tanh
Dense Layer	Fully Connected	32 units	ReLU
Output Layer	Fully Connected	Demand prediction output	Linear

The architecture of the proposed LSTM-based forecasting model is depicted in Figure 3. which illustrates the conceptual flow of information through the proposed LSTM-based forecasting model. Historical electricity demand measurements, together with meteorological and temporal variables, are processed through successive LSTM layers to capture temporal dependencies before generating short-term electricity demand forecasts.

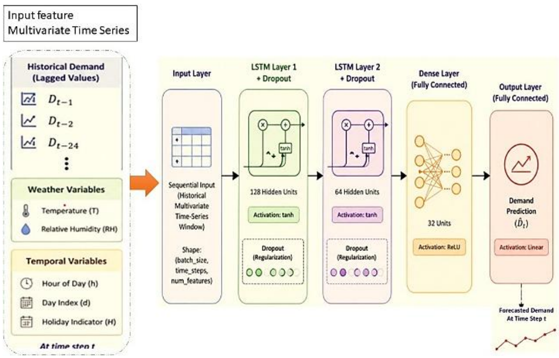

Fig. 3. Proposed LSTM-Based Electricity Demand Forecasting Architecture.

6.2.4 Proposed LSTM Model's Conceptual Training Setup

The proposed Long Short-Term Memory (LSTM) network is conceptually designed to support short-term electricity demand forecasting within the smart grid framework. The model is intended to capture temporal dependencies in electricity consumption data using sequential learning mechanisms.

Within the proposed framework, the LSTM model is assumed to operate on historical multivariate time-series data obtained from smart meters, meteorological systems, and temporal indicators. The model design incorporates standard practices commonly used in time-series forecasting, including sequential input structuring and time-dependent feature representation.

The actual training process, optimization strategy, and empirical validation are not performed in this study and are reserved for future implementation and evaluation within real operational datasets.

6.2.5 Forecasted Performance Characteristics

Within the proposed architecture, the LSTM model is intended to provide short-term electricity demand forecasting capabilities by learning nonlinear temporal relationships in historical consumption patterns and external influencing factors such as weather and calendar effects.

Existing literature has demonstrated that LSTM-based models are effective for load forecasting due to their ability to capture long-term dependencies in sequential data. Accordingly, the LSTM model is included in the proposed framework as a suitable deep learning component for demand prediction tasks [10].

Quantitative performance evaluation metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and related indicators are not computed in this study. These are intended for future empirical validation using Ghana-specific electricity demand datasets.

6.3 Model 3: Grid Anomaly Detection: Isolation Forest

6.3.1 Model Rationale

Grid environments in Ghana are subject to operational disturbances such as voltage fluctuations, frequency instability, transformer overload, sensor malfunctions, and non-technical losses. In many real-world power systems, labelled anomaly datasets are scarce, particularly in developing-country contexts, making supervised anomaly detection approaches difficult to apply.

To address this limitation, the proposed framework adopts the Isolation Forest algorithm, an unsupervised learning method introduced by Liu et al. [11]. The model is based on the principle that anomalous observations are more susceptible to isolation through random partitioning of the feature space compared to normal data points.

6.3.2 Proposed Configuration and Feature Space

Within the proposed framework, the Isolation Forest model is conceptually designed to analyse multivariate measurements collected from smart grid monitoring systems in order to identify abnormal operating conditions. The suggested feature vector is as follows:

$$X = \left[V_{\{dev\}}, F_{\{dev\}}, \mathrm{DFR}, \mathrm{PF}, \Delta L, \mathrm{TL}\right] \tag{4}$$

where ($V_{\{dev\}}$) denotes voltage deviation, $F_{\{dev\}}$ represents frequency deviation, (DFR) denotes the demand-to-forecast ratio, (PF) represents power factor, (L) denotes sudden load variation, and (TL) represents transformer loading.

These variables are selected based on their relevance to grid instability, equipment stress, and operational inefficiencies, including potential non-technical losses. Within the conceptual framework, these features provide a basis for identifying abnormal operational patterns in smart grid environments.

6.3.3 Expected Detection Characteristics

Within the proposed architecture, the Isolation Forest model is intended to support unsupervised anomaly detection for monitoring abnormal grid behaviour, sensor inconsistencies, and unusual demand patterns. The model is included due to its suitability for high-dimensional, unlabelled operational data commonly encountered in power systems.

Performance evaluation and quantitative benchmarking of anomaly detection effectiveness are not conducted in this study. These are reserved for future empirical implementation using operational data from Ghana's electricity infrastructure.

6.4 Model Comparison and Selection Summary

The three machine learning models presented in this framework are selected based on their suitability for distinct functional roles within a smart grid environment.

Random Forest is conceptually adopted for solar photovoltaic power forecasting due to its effectiveness in modeling nonlinear relationships in environmental data. Long Short-Term Memory (LSTM) networks are included for electricity demand forecasting due to their ability to capture temporal dependencies in sequential data. Isolation Forest is selected for anomaly detection because of its suitability for unsupervised identification of abnormal patterns in high-dimensional operational datasets.

Collectively, these models form a conceptual analytics layer within the proposed big data-driven smart grid architecture for renewable energy integration and operational intelligence in Ghana.

7. GHANA-SPECIFIC IMPLEMENTATION CONTEXT

7.1. Regulatory and Policy Landscape

The renewable energy and digital transformation landscape in Ghana provides an enabling policy environment for the development of intelligent smart grid and big data analytics frameworks. The Renewable Energy Act, 2011 (Act 832), together with the Renewable Energy Master Plan (2019–2030), promotes increased renewable energy integration, grid modernization, and sustainable energy development.

In addition, national institutions such as the Energy Commission, the National Information Technology Agency (NITA), and the Ghana-India Kofi Annan Centre of Excellence in ICT (AITI-KACE) contribute to the development of digital infrastructure and ICT capacity. These institutions provide an enabling ecosystem that supports the conceptual deployment of advanced data-driven energy management frameworks within the Ghanaian power sector.

7.2 Existing Data Infrastructure

Ghana's power sector generates multiple streams of operational data that are relevant to the proposed conceptual framework. For example, the Ghana Grid Company (GRIDCo) operates SCADA systems that provide transmission-level measurements of grid conditions, while the Electricity Company of Ghana (ECG) is progressively deploying Advanced Metering Infrastructure (AMI) for consumer-level electricity monitoring.

In addition, meteorological and environmental data are available from the Ghana Meteorological Agency (GMet) and external sources such as the NASA POWER API. These datasets provide complementary information on solar irradiance, temperature, humidity, and atmospheric conditions, which are relevant for renewable energy forecasting.

Within the proposed framework, these data sources are considered as potential inputs for big data analytics and machine learning-based decision-support systems. However, their use in this study is limited to conceptual illustration rather than empirical model development or validation.

7.3 Constraints and Mitigation Strategies

The implementation of smart grid analytics frameworks in Ghana is associated with several technical, infrastructural, and organizational constraints. Table 4 summarizes the principal technical, infrastructural, and organizational challenges that may affect implementation of the proposed framework in Ghana. The accompanying mitigation strategies demonstrate how these challenges could be addressed through phased deployment, improved data management, cybersecurity measures, and institutional collaboration.

Table 4. Ghana-Specific Constraints and Proposed Mitigation Approaches.

Constraint	Relative Impact	Proposed Mitigation Strategy	Relevant Stakeholders
Limited network connectivity in some grid regions	High	Edge-computing support and delayed synchronization mechanisms	GRIDCo / Telecom Providers
Limited smart-meter deployment coverage	High	Use of aggregate feeder-level demand estimation in low-AMI regions	ECG / Energy Commission
Harmattan-related atmospheric variability	Medium	Seasonal forecasting adaptation using aerosol and weather indicators	Energy Research Institutions
Sensor noise and missing operational data	Medium	Data-cleaning, imputation, and anomaly-filtering pipelines	Utility Data Teams
Shortage of Big Data engineering expertise	High	Academic–industry partnerships and specialized technical training programs	Universities / ICT Institutions
Cybersecurity risks within smart-grid environments	High	Network segmentation, encryption, and access-control mechanisms	Cybersecurity Agencies / Utilities
Infrastructure and deployment costs	Medium	Phased implementation and hybrid cloud-support strategies	Energy Commission / Government

7.4 Conceptual Implementation Roadmap

This study proposes a phased conceptual roadmap for the potential adoption of the big data-driven smart grid framework within Ghana's electricity sector. The roadmap is intended as a strategic guide rather than a fixed implementation schedule, and its actual execution would depend on institutional readiness, funding availability, and regulatory approval. Table 5 presents the proposed phased implementation roadmap for the conceptual framework. The roadmap illustrates how the architecture could evolve from pilot deployment to nationwide implementation, while emphasizing that the phases represent a strategic guideline rather than an executed implementation plan.

The roadmap is intended to illustrate how the proposed framework could evolve under ideal conditions. It does not represent an executed deployment plan.

Table 5. Proposed Phased Implementation Roadmap.

Phase	Timeline	Proposed Scope	Key Objectives
Phase 1: Foundation	Year 1	Pilot deployment within selected operational districts	Establish a distributed data ingestion and storage infrastructure
Phase 2: Expansion	Years 2–3	Regional-scale expansion of smart-grid analytics capabilities	Introduce forecasting and anomaly-detection modules
Phase 3: National Integration	Years 4–5	Broader national-grid integration and interoperability	Expand decision-support and regional coordination capabilities
Phase 4: Optimization	Long-Term	Continuous analytics improvement and operational refinement	Support AI-assisted grid management and adaptive forecasting

7.5 Availability of Electricity Demand Data (Illustrative Context)

Publicly available electricity demand data from Ghana, including historical peak demand records from GRIDCo, provide useful insights into national load variation patterns. These datasets capture temporal variations in electricity consumption and highlight the importance of forecasting in grid management.

Within this study, such datasets are used solely as illustrative examples to demonstrate the relevance of the proposed forecasting components within the conceptual framework. They are not used for model training, validation, or performance evaluation

Figures 4 and 5 provide an illustrative overview of historical electricity demand patterns in Ghana using publicly available GRIDCo data. Figure 4 shows the daily variation in peak electricity demand, highlighting short-term fluctuations that motivate the need for accurate demand forecasting. Figure 5 presents the monthly average peak demand, revealing seasonal variations in electricity consumption. Together, these figures demonstrate the practical relevance of forecasting within the proposed framework while serving solely as illustrative examples rather than inputs for model training or validation [20].

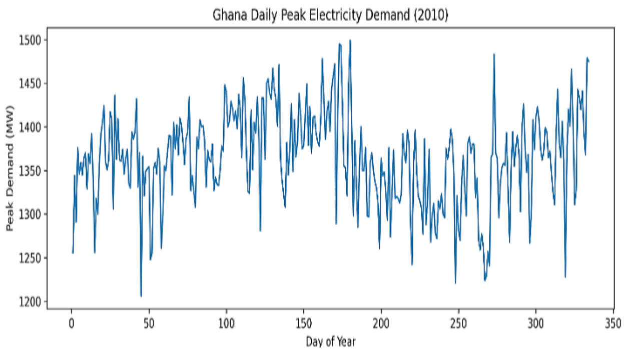

Fig. 4. Profile of peak electricity demand per day in Ghana for the year 2010 as provided by GRIDCo.

This graph depicts changes in electricity demand over time and demonstrates the importance of short-term electricity demand prediction in smart grids.

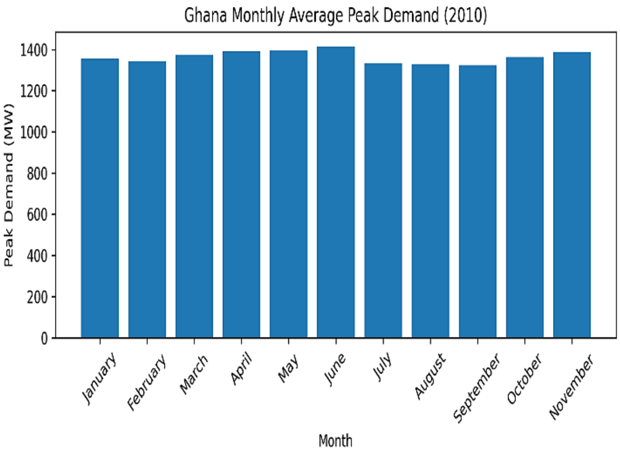

Fig. 5. Monthly average peak electricity demand in Ghana during 2010.

The observed seasonal variations in electricity consumption justify the adoption of advanced forecasting techniques such as Long Short-Term Memory (LSTM) networks within the proposed framework

7.6 Summary of Ghana Context

Overall, the Ghanaian energy sector presents both opportunities and challenges for the adoption of advanced data-driven smart grid systems. While policy frameworks and data availability provide a strong foundation, limitations in infrastructure, expertise, and system integration necessitate a phased and carefully designed approach.

The proposed conceptual framework is therefore positioned as a strategic model to guide future development of scalable, intelligent, and data-driven energy management systems tailored to the Ghanaian context.

8. CHALLENGES, OPPORTUNITIES, AND POLICY RECOMMENDATIONS

8.1 Technical Challenges

In addition to the practical issues outlined above, the proposed approach is subject to a number of additional technical challenges that are common for all kinds of Big Data applications in smart grid settings. First and foremost, there is the issue of data drift, which means that over time, the statistical properties of datasets used for operation will slowly start changing due to the influence of such factors as sensor aging, degradation of the power distribution network, climate changes, expansion of the grid, and changes to the renewable energy generation infrastructure.

The concept of the framework involves the introduction of some kind of mechanisms for periodic model updates and performance monitoring that are expected to help mitigate this problem. Another technical challenge refers to interoperability between different utility systems and communication protocols, since smart-grid environments typically involve both older operational infrastructure and relatively new technological solutions.

8.2 Ethical and Governance Considerations

High-frequency smart meter and grid operations data could be sensitive in nature and pose risks in terms of identifying household occupancy behavior patterns, electricity consumption behavior, and business operations. For this reason, the proposed methodology adopts principles of data governance consistent with the provisions of Ghana's Data Protection Act, 2012 (Act 843) [21].

Data minimization, access control, anonymization practices, and authorization practices are key elements in the proposed methodology to minimize risks due to the collection of high volumes of operational data. Furthermore, feeder level aggregation is recommended to prevent unnecessary disclosure of individual customer level electricity consumption data to higher levels.
Another important aspect of governance considered in this research is cybersecurity since smart grids have become highly vulnerable to cybersecurity attacks over networks.

8.3 Possible Opportunities and Benefits

Several operational and economic benefits may arise from the proposed framework in relation to Ghana's electricity industry. Accurate forecasts will enable more efficient generation-reserve management, better utilization of renewable energy sources, lower operational uncertainty, and better grid stability.

Additionally, the ability to perform real-time analytics and detection of anomalies may allow quicker detection of any operational inconsistencies, leading to more efficient resource utilization and maintenance operations. This framework is expected to benefit any ongoing efforts at smart grid development and aid Ghana in preparing for more renewables in regional electricity systems.

While quantitative economic benefits remain difficult to assess without an implementation-based study, previous research on smart grids has shown that more accurate forecasts and better decision-making support will significantly enhance operational and energy management efficiency.

8.4 Policy Recommendations

- The Energy Commission is advised to persistently encourage the development of common standards for smart-metering and renewable energy data that would be able to facilitate future smart grid interoperability and analysis capabilities.
- The Ghana Grid Company might explore the creation of specialized units focused on conducting big data analytics related to grid operations and renewable energy forecasting.
- Academic centers and institutions within Ghana should increase their focus on educating people in the domains of Big Data technologies, artificial intelligence, cybersecurity, and smart grid analytics, due to new skill demands.
- The Cyber Security Authority should proceed with formulating sector-specific cyber security policies for critical energy infrastructures and operational systems.
- Cooperation at the regional level could be achieved via the Economic Community of West African States and the West African Power Pool.

9. CONCLUSION

This study presents a conceptual Big Data architecture framework for renewable energy integration and smart grid decision-making within the Ghanaian electricity sector. The framework addresses the growing need for data-driven approaches to support forecasting, monitoring, and operational intelligence in modern power systems.

The proposed architecture integrates multiple layers, including data acquisition, streaming ingestion, distributed storage, data processing, machine learning analytics, and visualization. Open-source big data technologies such as Apache Kafka, Apache Spark, Hadoop Distributed File System (HDFS), and Apache Cassandra are conceptually incorporated due to their suitability for large-scale distributed data processing environments.

Three machine learning models are conceptually included to support key smart grid functions: Random Forest for solar photovoltaic forecasting, Long Short-Term Memory (LSTM) networks for electricity demand prediction, and Isolation Forest for anomaly detection. These models are selected based on findings from existing literature rather than empirical implementation within this study.

It is important to emphasize that this work does not present an implemented or experimentally validated system. Instead, it provides a structured conceptual blueprint intended to guide future research and practical deployment in smart grid environments.

The study further highlights key considerations for potential implementation in Ghana, including infrastructure readiness, cybersecurity, data governance, technical capacity, and regulatory support. Finally, the framework provides a foundation for future research involving empirical validation, prototype development, and advanced analytics techniques for smart grid systems.

REFERENCES

[1] N. Höhne, M. J. Gidden, M. den Elzen, F. Hans, C. Fyson, A. Geiges, M. L. Jeffery, S. Gonzales-Zuñiga, S. Mooldijk, W. Hare, and J. Rogelj, "Wave of net zero emission targets opens window to meeting the Paris Agreement," *Nature Climate Change*, vol. 11, pp. 820–822, 2021, doi: 10.1038/s41558-021-01142-2.

[2] Energy Commission of Ghana, *Ghana Integrated Power System Master Plan*, vols. 1–3. Accra, Ghana: Energy Commission, 2018.

[3] F. Nwaiwu, "Digitalisation and sustainable energy transitions in Africa: Assessing the impact of policy and regulatory environments on the energy sector in Nigeria and South Africa," *Energy, Sustainability and Society*, vol. 11, Art. no. 48, 2021, doi: 10.1186/s13705-021-00325-1.

[4] C. M. Mulevu, G. O. Okeyo, and J. M. Wafula, "Data stream mining techniques for real-time monitoring and control of smart power grids in Kenya: Challenges and opportunities," *Discover Internet of Things*, vol. 5, Art. no. 51, 2025, doi: 10.1007/s43926-025-00146-0.

[5] G. M. Morara and E. M. Ondieki, "Leveraging artificial intelligence for renewable energy forecasting in Kenya: Implications for sustainable development," *International Journal of Energy Sector Management*, pp. 1–17, Feb. 2026, doi: 10.1108/IJESM-08-2025-0036.

[6] International Renewable Energy Agency (IRENA), *World Energy Transitions Outlook 2022: 1.5°C Pathway*. Abu Dhabi, United Arab Emirates: IRENA, 2022.

[7] H. Daki, A. El Hannani, A. Aqqal, A. Haidine, and A. Dahbi, "Big data management in smart grid: Concepts, requirements and implementation," *Journal of Big Data*, vol. 4, Art. no. 13, 2017, doi: 10.1186/s40537-017-0070-y.

[8] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[10] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841–851, 2019, doi: 10.1109/TSG.2017.2753802.

[11] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE International Conference on Data Mining (ICDM)*, Pisa, Italy, Dec. 2008, pp. 413–422, doi: 10.1109/ICDM.2008.17.

[12] Ishwarappa and J. Anuradha, "A brief introduction on Big Data 5Vs characteristics and Hadoop technology," *Procedia Computer Science*, vol. 48, pp. 319–324, 2015, doi: 10.1016/j.procs.2015.04.188.

[13] H. A. E.-S. Ali, M. H. Alham, and D. K. Ibrahim, "Big data resolving using Apache Spark for load forecasting and demand response in smart grid: A case study of Low Carbon London Project," *Journal of Big Data*, vol. 11, Art. no. 59, 2024, doi: 10.1186/s40537-024-00909-6.

[14] E. N. Kumi, "The electricity situation in Ghana: Challenges and opportunities," CGD Policy Paper No. 109, Center for Global Development, Washington, DC, USA, Sep. 2017. [Online]. Available: https://www.cgdev.org/publication/electricity-situation-ghana-challenges-and-opportunities. Accessed: Jun. 27, 2026.

[15] N. Marz and J. Warren, *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*. Shelter Island, NY, USA: Manning Publications, 2015.

[16] J. Im, J. Lee, S. Lee, and H.-Y. Kwon, "Data pipeline for real-time energy consumption data management and prediction," *Frontiers in Big Data*, vol. 7, Art. no. 1308236, 2024, doi: 10.3389/fdata.2024.1308236.

[17] A. Ashabi, M. Mostafa, A. Hryshchenko, K. Bruton, and D. T. J. O'Sullivan, "Design and utilization of a decision support tool to advance energy efficiency in industries," *Energy*, vol. 308, Art. no. 132975, Nov. 2024, doi: 10.1016/j.energy.2024.132975.

[18] M. S. Adaramola, "Estimating global solar radiation using common meteorological data in Akure, Nigeria," *Renewable Energy*, vol. 47, pp. 38–44, 2012, doi: 10.1016/j.renene.2012.04.005.

[19] V. Raj, S.-Q. Dotse, M. Sathyajith, M. I. Petra, and H. Yassin, "Ensemble machine learning for predicting the power output from different solar photovoltaic systems," *Energies*, vol. 16, no. 2, Art. no. 671, 2023, doi: 10.3390/en16020671.

[20] Ghana Grid Company Ltd. (GRIDCo), "Ghana—National daily peak demand for electricity, 2009–2010," *EnergyData.info*, World Bank Group. [Online]. Available: https://energydata.info/en/dataset/ghana-national-daily-peak-demand-electricity-2009-2010. Accessed: Jun. 27, 2026.

[21] T. Hoffmann, "Data Protection Act(ion) – Report on the Law of Data Disclosure in Ghana," *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4037928.