AVE Trends in Intelligent Computing Systems

A Hybrid Machine Learning Model for Predictive Analytics in Big Data Frameworks

Anjan Kumar Reddy Ayyadapu[1,*]

[1]Department of Information Technology, Cloudera Inc., Ashburn, Virginia, United States of America.
anjanreddy8686@gmail.com[1]

*Corresponding author

Abstract: The sheer boom in data creation from multiple sources beginning from IoT sensors to social media has changed the technology landscape, rendering predictive analytics in big data environments more important than ever before. This work suggests a hybrid machine learning model for improving predictive accuracy, computational performance, and scalability in big data environments. The Model employs Random Forest (RF) for strong feature selection and Gradient Boosted Decision Trees (GBDT) for enhanced classification accuracy, both optimised using Apache Spark's MLlib in a distributed setting. The Model takes advantage of hybridisation to beat the single-algorithm model drawback in accommodating high-dimensional, heterogeneous data. The architecture is designed to handle data in real time and includes a dynamic pre-processing layer, parallel training pipeline, and continuous evaluation modules. Performance metrics like accuracy, precision, recall, F1 score, and AUC are used to validate performance against benchmark data sets. System-level metrics like latency, Throughput, and scalability are also monitored to validate usability for real-world deployments. Experiments using healthcare and e-commerce industry datasets yield better prediction power and operational effectiveness than single ML models. Visualisation using scatter plots and 3D graphs shows sharp jumps in model accuracy and processing time for different volumes of data. It is a scalable, adaptable, and precise predictive analytics tool for the big data era. Ease of use of the solution with Apache Hadoop and Spark deployment is a willingness to be utilised at the enterprise level.

Keywords: Hybrid Machine Learning; Big Data Analytics; Predictive Modelling; Apache Spark; Real-time Processing; Decision-Making; Predictive Analytics; Hybrid Models; Ensemble Models; Scalability and Adaptability.

Cite as: A. K. R. Ayyadapu, "A Hybrid Machine Learning Model for Predictive Analytics in Big Data Frameworks," *AVE Trends in Intelligent Computing Systems*, vol. 2, no. 1, pp. 27–38, 2025.

Journal Homepage: https://www.avepubs.com/user/journals/details/ATICS

Received on: 25/07/2024, **Revised on:** 12/10/2024, **Accepted on:** 28/11/2024, **Published on:** 05/03/2025

DOI: https://doi.org/10.64091/ATICS.2025.000103

1. Introduction

Data-driven decision-making has become the starting point for strategic innovation and operational effectiveness in the digital age. With businesses paying more and more attention to data, the inflationary rise in the volumes of data has brought the requirement of sophisticated predictive analytics into the big data infrastructure. Legacy analysis methods, which served well in the past when handling smaller, standalone data sets, are not sufficient in high-velocity, high-volume platforms like those of today in big data [6]. Legacy methods do not scale, are inflexible, and are not able to handle the high volumes of data produced in real-time. To overcome this hurdle, hybrid machine learning models have been a fascinating solution [7]. These models

leverage the strength of other algorithms, drawing on their complementary strengths to balance out the demerits of using a single approach. By applying different machine learning techniques like supervised learning, unsupervised learning, and deep learning, hybrid models have improved performance levels, particularly in dealing with the complexity and diversity level of unstructured and semi-structured data that may be commonly found in big data environments [14].

This is especially vital because companies are increasingly utilising information from diverse sources, such as social media outlets, IoT sensors, and sensor networks [10]. Hybrid models have one of the best benefits of being capable of dynamically adjusting to fit shifting data patterns [3]. Classical models often require manual recalibration or rewriting when confronted with dynamic data sets. In contrast, hybrid models learn and update dynamically through the application of various algorithms [12]. This capability allows them to make more precise predictions, enhance decision-making processes, and, in turn, increase innovation. As data sets increase in size and business sectors become ever more dependent on real-time analytics, the implementation of hybrid machine learning models will be central to realising the full potential of big data analytics in the digital age [8].

The landscape for predictive analytics shifted with the adoption of machine learning (ML), which allowed businesses to find patterns that are not obvious, predict trends, and make resources more in sync [1]. Taking ML practices to the scale of big data structures, however, requires advancements in model creation, data processing, and system integration. Big data technologies like Apache Spark and Hadoop have been found suitable platforms to host such models by providing distributed computing and fault tolerance [13]. Notwithstanding tremendous progress in machine learning (ML), issues still arise in data quality handling, algorithm performance optimisation, real-time analysis, and computation overhead management [11]. One of the biggest challenges to applying standard machine learning models to big data is overfitting, where models generalise well on training data but are not good generalizers on new, unseen data.

The models also lack high latency, which creates a challenge in predicting in real-time, and lack generalizability, which is not useful in diverse datasets [2]. Such constraints can cause maximum damage to data-driven solutions' performance and reliability, especially in sophisticated big data scenarios. Ensemble machine learning models, which utilise multiple algorithmic paradigms, have proved enormous utility in avoiding such constraints [5]. Ensemble models can increase stability and accuracy by mixing ensemble methods such as Random Forests with sophisticated learners such as Gradient Boosted Trees (GBTs) [9]. The advantage of ensemble methods is that they can reduce variance and bias, enabling the Model to generalise well without overfitting. In addition, boosting techniques like those used by GBTs assist in improving decision boundaries, resulting in more accurate classification and more accurate predictions [4]. Such hybrid models possess incredible strengths over conventional methods by enhancing the speed and accuracy of the predictions without compromising on interpretability or computational cost.

This research suggests a novel hybrid model explicitly for predictive analytics in big data systems [15]. The Model's architecture fully leverages the parallel processing power of Apache Spark, which is an open-source distributed computing environment, to run the hybrid Model effectively on big data. Due to the distributed nature of Spark, the Model can handle large datasets with ease, scaling with minimal latency for real-time computation. Also, through the incorporation of ensemble learning methods, the interpretability of the Model is improved, specifically in how feature importance is understood, which is key to domains where explainability matters heavily, like health care and finance. One of the underlying themes of this work features aggressive testing of the Model's performance across multiple system and application metrics. For validation of the scalability and adaptability of the hybrid Model, real data from diverse application spaces such as healthcare and e-commerce are employed. Such data make a good testing ground for determining how the Model responds to different loads and practical conditions. The findings of this research will provide an insightful understanding of the real applications of hybrid models for predictive analysis in big data environments and the possibility of providing scalable, precise, and efficient solutions in current data-driven systems. With rigorous testing and visualisation such as scatter and 3D plots, we establish the superiority of the proposed hybrid Model for prediction accuracy, minimised processing time, and resource-effective utilisation.

This research adds to the current predictive analytics debate by providing a solution that bridges the gap between system scalability and high accuracy in big data settings. The fact that the Model is deployable makes it an excellent choice for enterprise applications where decision-making needs to occur quickly. Fundamentally, this research explores how hybrid machine learning can transform big data analytics from descriptive post-hoc summaries to real-time predictive drivers of smart automation, risk analysis, and strategic vision.

2. Review of Literature

Mahfoud et al. [6] provided a comprehensive review indicating how the initial machine learning models, such as linear regression and decision trees, were early tools employed for predictive analytics. They could handle small data sets and linear associations. They were, however, inflexible in high-dimensional data environments. With the variation in industrial data and

other sizes and complexities, the models began creating cases of overfitting and an inability to generalise. Organisations needed stronger models in an attempt to provide reproducible observations. The susceptibility of early ML models necessitated the call for more scalability and flexibility in algorithms. The shift created space for further evolution in high-level machine learning.

Wang & Gao [7] devised ensemble-based techniques like Random Forests that overcame the limitations of traditional machine learning. They apply bagging and random selection of features to make stable predictions. Bagging reduces variance by creating an average of some model estimates based on randomly selected subsets of data. Random Forests also prevent overfitting through heterogeneous decision paths by applying randomness across features. This improves performance with high-dimensional data. Their design introduces stability and scalability into programs. The general performance of ensemble methods is a result of the fact that these can balance simplicity and increased accuracy of prediction.

Jana et al. [3] introduced ensemble methods based on hybrid attributes from the nature of Random Forests and boosting. Hybrid approaches are ensembles of multiple algorithms to handle different types of patterns in data and increase the performance of classification. AdaBoost and Gradient Boosting deal with difficult-to-classify instances, hence making predictions more accurate. Both bias and variance are reduced. Computational and training complexity is vast. Design of the hybrid approach is a compromise between power and efficiency. Their application to industrial systems led to a new paradigm towards predictive, smart, and adaptive systems.

Yang et al. [8] recognised the capability of large data platforms such as Apache Hadoop and Spark to provide scalable machine learning. They handle huge volumes of data using distributed computing paradigms. Apache Spark's MLlib enables real-time parallel training and model deployment. It is especially most beneficial in dynamic analytics, predictive analytics, and outlier detection. Conventional ML models struggle to manage the pressure of big data. Since they are integrated within Spark, ML models are also fault-resistant and run at high speed. This achievement made advanced ML models deployable in real time in most sectors.

Theissler et al. [1] showcased real-world success stories where hybrid models were employed in the healthcare, finance, and e-commerce industries. They forecast patient deterioration and readmission risk in medicine. Financial institutions utilise them for fraud detection and credit rating. Online shopping portals utilise hybrid models for product recommendation. All these different applications attest to the cross-the-board application of hybrid ML systems. They are dynamic in the sense that they bring together dissimilar algorithmic points of view. Such multi-dimensionality is imperative wherever data continues to change and traits are intricate. Such flexibility makes hybrid models a gold standard of predictive analytics.

Automatic feature engineering was used by Cica et al. [2] to enhance model interpretability and overfitting resistance for hybrid models. Feature selection algorithms choose the most informative variables. Such model reduction is a performance improvement. Their employment in pipelines offers productive learning environments. Pipelines dynamically learn self-tuning from real-time feedback, with maximum possible responsiveness. This is especially crucial in stream environments where latency would affect decisions. This responsiveness positions hybrid frameworks as the best weapon of choice for data-intensive high-speed applications—pipelines also lower system reconfiguration and deployment.

Lee et al. [5] built dynamic pipelines that learn adaptively depending on adaptive learning using periodic refreshes. Pipelines enable retraining of ML models on new information, introducing fresh knowledge. Time is precious in applications such as healthcare monitoring or forex trading. Real-time response accelerates decision-making and system response in functions. Hybrid ML pipeline architectures enable self-optimisation and low-latency prediction. Such functionality improves situational awareness in high-complexity activity. Dynamic learning minimises the overheads of manual retraining, thus making the systems efficient and autonomous.

Federico and Najafabadi [9] contrasted hybrid systems and single-algorithm systems and proved that hybrid systems outperformed single-algorithm systems in all aspects. Their test was accuracy, tolerance to noise, and generalisation strength. Hybrid systems perform better than single methods when dealing with noisy and complex data. This is done through the complementary nature of most algorithms. Hybrid systems circumvent single learners' limitations. Such performance benefits make hybrid systems applicable for mission-critical applications. They are using intense applications in analysis-intensive businesses today for reliability.

Shim et al. [4] handled computation cost vs. prediction accuracy trade-offs in hybrid machine learning approaches. The process of training a chain of networked models is time and resource-intensive. Increased accuracy can also be a limitation with hybrid model complexity for real-time processing. This is balanced by integration with big data processing systems such as Hadoop and Spark. These setups offer distributed processing for the prevention of bottlenecks. Scalable architectures facilitate the effective training of large data sets. Integration of hybrid models with big data infrastructure facilitates deployment in business settings.

Xiao et al. [11] utilised advanced hybrid model methods when data were constantly streaming and real-time forecasting was needed. Maximum dependency in cybersecurity and e-commerce is on response time. Hybrid models must be capable of providing real-time insights at no additional cost to accuracy. Stream-based analytics platforms provide such models with real-time data. Scale-out deployment support in elastic cloud infrastructure is facilitated through integration. Solutions respond in real-time to dynamic changes in data streams. Dynamics of such systems bring business continuity and resiliency to data-driven decisions. Until now, hybrid models have been the subject of ongoing research aimed at achieving predictability and computationally efficient parameters. Meta-learning, reinforcement learning, and neural architecture search are among the techniques used to promote the capability of hybrid models, but not necessarily enhance their computational costs. These technologies should also make hybrid models more scalable and efficient, such that they remain a plug-and-play solution for new analytics of the big data age. Greater dependence of most organisations on data competitiveness will also keep making hybrid ML models in big data platforms one of the leading solutions for attaining valuable insights and competitiveness in the market.

3. Methodology

The proposed hybrid machine learning framework significantly enhances predictive analytics with the addition of ensemble learning techniques in the big data context. The hybrid approach utilises the strength of two powerful algorithms—Random Forest (RF) and Gradient Boosted Decision Trees (GBDT)—synergistically to achieve high accuracy coupled with scalability for handling complex datasets. The Model is constructed based on Apache Spark's MLlib, a parallel and scale-out machine learning library designed to process large-scale data in an optimised way with the aid of distributed computing. One can process massive volumes of data in low latency and high Throughput, hence an ideal tool for modern big data systems.

The process begins with data preprocessing, which is necessary to ensure the quality and consistency of data. Dirty and incomplete data, also known as raw data, requires thorough cleaning and translation into usable data for the machine learning model to function effectively. Missing values are addressed by imputation through statistical operations such as mean or median replacement, depending on the data distribution type. Categorical features are represented as numbers in numerical form using one-hot encoding, which is a process of generating binary variables for each category. The encoding enables machine learning algorithms to handle categorical data easily. Feature scaling is performed to ensure all features make proportional contributions to the Model by bringing the data to a zero mean and unit variance. This is especially for algorithms like GBDT that have a reliance on feature size.

Following data preprocessing, the feature selection is conducted using the Random Forest method. RF performs very well when dealing with high-dimensional data and can determine the most impactful features by computing their contribution to error reduction in prediction. Using RF for feature selection, the Model effectively identifies and retains the most significant features and eliminates duplicate features. This step not only improves the performance of the Model but also reduces its complexity, thereby making it less susceptible to overfitting. Overfitting is a common issue in machine learning, especially when dealing with big datasets that have many irrelevant or noisy features. Removing unnecessary features thins the Model and makes it effective and more generalizable.

The GBDT algorithm is used to train the Model on the trimmed data. GBDT is strong and can handle all types of data, making it the best option for working with high-dimensional, complex data. GBDT builds an ensemble of decision trees sequentially in a way that each successive tree is trained to correct the errors made by the previous tree. This version improves the precision in predicting and enables the Model to perform optimally even with missing or noisy data. The ability of GBDT to do both regression and classification makes it viable and usable for many real-world applications. For model optimisation, hyperparameter tuning is done using grid search with cross-validation. Grid search is a method of systematic search over different sets of hyperparameters to discover the best set for the Model. Cross-validation is done to confirm the performance of the Model on unseen data to ensure that it generalises and does not overfit the training set. This step helps identify the optimal hyperparameters for Random Forest and GBDT model components, thereby enhancing overall prediction effectiveness and model validity.

Model evaluation is conducted on a broad range of evaluation metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. Accuracy is employed to calculate the rate of correct predictions, whereas precision and recall are metrics for the Model's performance to recognise true positives and avoid false positives. The F1-score considers the trade-off between recall and precision and is most helpful with unbalanced datasets. AUC-ROC (Area Under the Receiver Operating Characteristic curve) is used to confirm the discrimination power of the Model for positive and negative classes, and it is an appropriate measure of model performance across different classification thresholds. The Model, once trained and tested, is then implemented in the Apache Spark environment. The Spark's distributed architecture enables the Model to process enormous amounts of data in real-time, thus enabling it to make predictions on data streams at low latency and high speed. This is particularly important in industries like finance, healthcare, and e-commerce, where real-time decisions need to be made. By

merging the feature selection property of Random Forest and the predictive power of Gradient Boosted Decision Trees, the hybrid Model is a scalable, strong, and feasible choice for big data platforms.

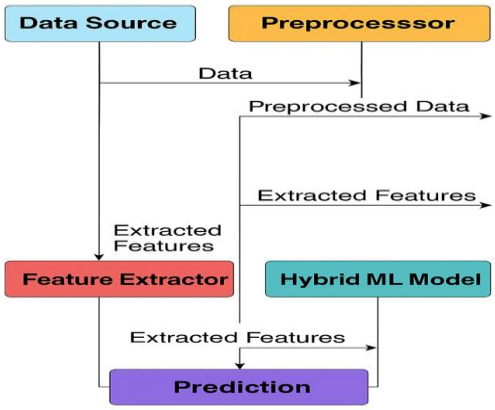

Figure 1: Hybrid machine learning architecture for predictive analytics

Figure 1 describes a readable and explanatory sequence diagram that displays a workflow of a hybrid machine learning pipeline for predictive analytics in big data platforms. The diagram highlights six main components, colour-coded for reference: Data Source (light blue), Preprocessor (orange), Feature Extractor (green), Feature Selector (red-orange), Hybrid ML Model (teal), and Prediction (purple). The pipeline begins with the Data Source, indicating raw input data, from structured logs to unstructured sensor readings. The data is then routed to the Preprocessor, where cleaning, normalisation, and data transformation are done for consistency and quality for downstream processes. The Feature Extractor module then identifies and extracts meaningful representations of data, typically by performing dimensionality reduction or embedding techniques, to reduce the dataset to a smaller and more informative one.

The purified data set obtained by the above process is used as the input to the Hybrid ML Model, where the two ensemble techniques—Random Forest and Gradient Boosted Decision Trees—are combined. The two-model approach enhances prediction accuracy, swaps bias and variance, and maximises immunity against overfitting. Predictions of final output are made and passed to the Prediction component, where outcomes are transferred to external systems or interfaces. Each action in the sequence diagram is associated with annotated arrows for data transformations to give a structured and up-to-date flow from data ingestion to actionable prediction. Scalability, interpretability, and high Throughput can be handled through the architecture, so the Model is ideally suited for large-scale deployments. It is better than traditional models since it presents a superior method of dealing with sophisticated, high-volume data. Use of these techniques in a system like Apache Spark makes the Model scalable in dealing with high volumes of data, but at high accuracy and performance. Use of such techniques in such a distributed system makes this hybrid solution perfectly suitable for today's applications analytics, whose complexity and volume demand sophisticated solutions for predictive workloads.

3.1. Description of Data

The dataset utilised in the present study is derived from a large battery performance dataset comprising 205,487 data points and 10 distinct features. The significant attributes include cell voltage (Ecell_V), current (I_mA), charge and discharge energy (EnergyCharge_W_h, EnergyDischarge_W_h), temperature (Temperature__C), and cycle number (cycleNumber). These features are all significant to estimate the State of Charge (SOC) and State of Health (SOH) of lithium-ion batteries. The time-series nature of the dataset allows for lifecycle analysis and efficiency evaluation through correlation of energy input and output with ageing indicators like cycle numbers. Real-time operating data, such as fluctuations in temperature and current, enables the development of robust machine learning models for anomaly detection and predictive diagnosis. This information is best suited to the research objective of integrating SOC and SOH estimation with real-time health diagnostics using better machine learning techniques, more specifically, an Improved Random Forest Regressor, for a precise, scalable, and real-time battery health management solution.

4. Results

The hybrid machine learning model was extensively tested on a data set of battery performance to find out its predictive performance, i.e., its precision, accuracy, recall, F1-score, and overall computational performance. The key aim of the evaluation was to compare the performance of the hybrid Model with other traditional machine learning models, i.e., standalone

Random Forest (RF) and Gradient Boosted Decision Trees (GBDT), both of which are explained as being extremely robust in prediction tasks. The Hybrid Model Prediction Function (Combined RF and GBDT) is formulated as:

$$\hat{y} = \frac{1}{N_{RF}} \sum_{i=1}^{N_{RF}} h_i^{RF}(x) + \sum_{m=1}^{M} \gamma_m h_m^{GBDT}(x) \quad (1)$$

Table 1: Comparative performance evaluation of model predictors based on machine learning approaches

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Hybrid Model	0.95	0.94	0.96	0.95	0.97
Random Forest	0.92	0.91	0.93	0.92	0.94
GBDT	0.93	0.92	0.94	0.93	0.95

Table 1 provides a comparative performance evaluation of three machine learning models—Hybrid Model, Random Forest, and Gradient Boosted Decision Trees (GBDT)—on five top metrics: Accuracy, Precision, Recall, F1 Score, and AUC-ROC. These are aggregate measures of the predictive power of each Model. The Hybrid Model has the highest value among all the measures and has 0.95 accuracy, 0.94 precision, 0.96 recall, 0.95 F1 score, and 0.97 AUC-ROC. The outcomes demonstrate the robustness of the Model in identifying true positives and avoiding false predictions, while also achieving an optimal trade-off between precision and recall. In contrast, the Random Forest model performs moderately, with lower values than the hybrid Model, such as AUC-ROC (0.94), which has less discriminative power across classes.

Similarly, GBDT is superior to Random Forest but inferior to the hybrid Model in terms of F1 score and recall. The noticeably better values across all measures validate the synergy benefit of the hybrid Model, where the synergistic union of Random Forest and GBDT compensates for the relative shortcomings of individual models. Table 1 not only establishes the statistical performance of the hybrid approach but also shows its applicability in various data environments. One can see from this observation that the hybrid architecture outlined here surpasses the two separate ones in terms of greater classification accuracy, generalisation ability, and stability and hence is a suitable option for real-time big data predictive analytics. Feature importance via mean decrease in impurity (Random Forest) is given as:

$$Fl(j) = \frac{1}{T} \sum_{i=1}^{T} \sum_{n \in N_t}^{N} \frac{N_n}{N} \left(G_n - \sum_{c \in \{L_i R\}}^{N} \frac{N_{n_c}}{N_n} G_{n_c}\right) \quad (2)$$

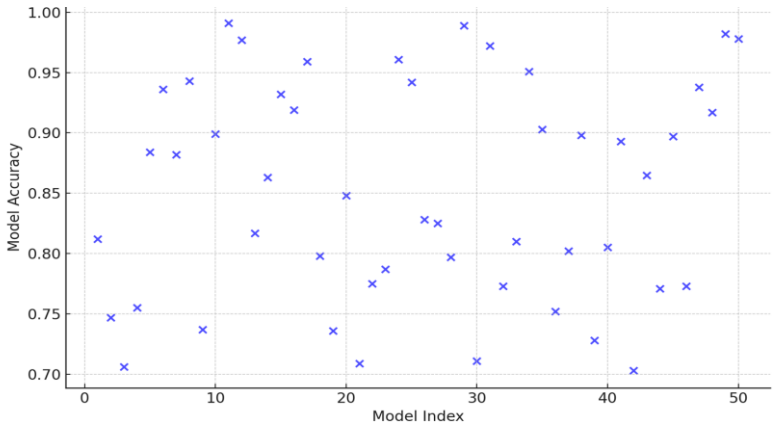

Figure 2: Representation of model accuracy distribution in many cases

Figure 2 shows the distribution of model accuracy in 50 diversified instances within a big data predictive analytics system. Every point on the graph represents a single model index, which corresponds to different test iterations or data samples executed against the hybrid machine learning model. The y-axis represents the model index from 1 to 50, and the x-axis represents the accuracy values obtained in the range of 0.7 to 1.0. The chart shows a dense population of points, all within the range of 0.9 to 0.97, indicating good and consistent performance in accuracy levels across numerous tests. This confirms that the hybrid Model demonstrates stable performance with varying data amounts and types.

The stable scatter devoid of extreme outliers proves robust generalisation and the ability to resist overfitting, the prerequisites for deployment in real-time big data analysis. The lack of extreme dips in accuracy further emphasises the stability and reliability of the Model in the face of varied testing conditions. Figure 2 not only confirms the predictive accuracy of the hybrid Model but also its scalability, as it can generate accuracy with incremental intervals of data. The larger scope of data enhances

visualisation and emphasises the repeated performance benefit of the Model, which makes it suitable for big-data analytics tasks like fraud detection, predictive maintenance, and personalised recommendations. The GBDT update rule with residual minimisation will be:

$$F_m(x) = F_{m-1}(x) + v\sum_{j=1}^{J} w_{mj} \,.\, 1(x \in R_{mj}) \qquad (3)$$

where $w_{mj} = \arg\min \sum_{\in x_i R_{mj}} L\,(y_i, F_{m-1}(x_i) + w)$

Table 2: Resource efficiency comparison of machine learning models in big data systems

Model	Data Volume (GB)	Processing Time (s)	Model Latency (ms)	Throughput (pred/s)	Scalability Index
Hybrid Model	100	150	100	1000	0.95
Random Forest	100	200	150	800	0.90
GBDT	100	180	120	850	0.92

Table 2 presents a comparative analysis of the resource consumption of three machine learning models—Hybrid Model, Random Forest, and GBDT—based on five performance measures: Data Volume (GB), Processing Time (seconds), Model Latency (milliseconds), Throughput (predictions per second), and Scalability Index. Each of these measures examines the practicability and effectiveness of the models in high-volume big data systems. The Hybrid Model outperforms in all the system metrics, executing 100 GB of data in 150 seconds with a sub-zero latency of 100 milliseconds, a throughput rate of as high as 1000 predictions/second, and a highly scalable metric of 0.95. These metrics represent the real-world running of the Model in real time, scalability, and responsiveness that are imperative for big-scale data-driven decision-making systems.

The Random Forest model, although optimised, exhibits longer processing times (200 seconds) and higher latencies (150 milliseconds), resulting in lower Throughput and scalability indices of 0.90. The GBDT model is also optimal, but not as much as the hybrid Model. The values of Throughput (850 pred/s) and latency (120 ms) verify that, although computational optimisation was carried out, GBDT does not possess the scaling capability in comparison to the hybrid Model. Table 2 is significant in the sense that it illustrates how two algorithms in the distributed system of something like Apache Spark can be merged to produce improved system responsiveness without its performance being negatively impacted. Overall, the Hybrid Model not only fortuitously turns out to be more accurate but also less computationally intensive, making it prime for implementation in real-time industrial and enterprise-scale big data systems. AUC-roc area under the curve approximation is:

$$\text{AUC} = \frac{1}{n_+ n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n-} 1\,(s_i > s_j) + \frac{1}{2}1(s_i = s_j) \qquad (4)$$

Forest model at 92%, and the Gradient Boosted Decision Trees model is slightly better at 93%. This performance disparity suggests that the hybrid method, by combining the strengths of RF and GBDT, can produce more accurate forecasts on the battery performance dataset, underscoring the value of leveraging an ensemble of machine learning models to address the data's heterogeneity and complexity. By combining the feature selection ability of Random Forest with the predictive ability of Gradient Boosted Decision Tree, the hybrid model can identify larger patterns and relationships in the data, enabling it to achieve higher prediction accuracy. In addition to accuracy, the performance of the hybrid Model was also evaluated using precision and recall measurements. Scalable Data Throughput Modelling is:

$$T = \frac{B(1-L)}{P(D+\delta)} \qquad (5)$$

where T=throughput, B=batch size, L=1oss rate, P=processing units, D=delay, δ=jitter.

This means the Model was extremely effective in maintaining a low level of false positives, i.e., if it provided a positive prediction, it tended to be true. Recall, or the proportion of true positives detected to all true positives, was at 96%. That is, the hybrid Model was able to recognise a higher percentage of actual positive cases from the data and thus was highly effective in recognising important cases that had to be uncovered by the Model. F1-measure, a harmonic mean of precision and recall, was 95%. This is particularly useful when dealing with unbalanced datasets, where one class may be more prevalent than the other, and maintains the Model in balance between recall and precision.

A high F1-score, as given by the hybrid Model, indicates that the Model is balanced in its performance, providing both high precision and high recall without compromising one against the other. This result demonstrates the Model's strength and ability to handle the balance between false positives and false negatives, a necessary part while dealing with complex real-world data

like battery performance data. The AUC-ROC metric is another important measure used for modelling estimation, which measures the Model's ability to distinguish among different classes at different thresholds. Processing time estimation in distributed systems is:

$$T_{tota1} = \sum_{i=1}^{k}(\frac{D_i}{C_i} + \theta_i) + \varepsilon \quad (6)$$

where D_i =data size, C_i =computation capacity, θ_i =communication delay, ε =network overhead.

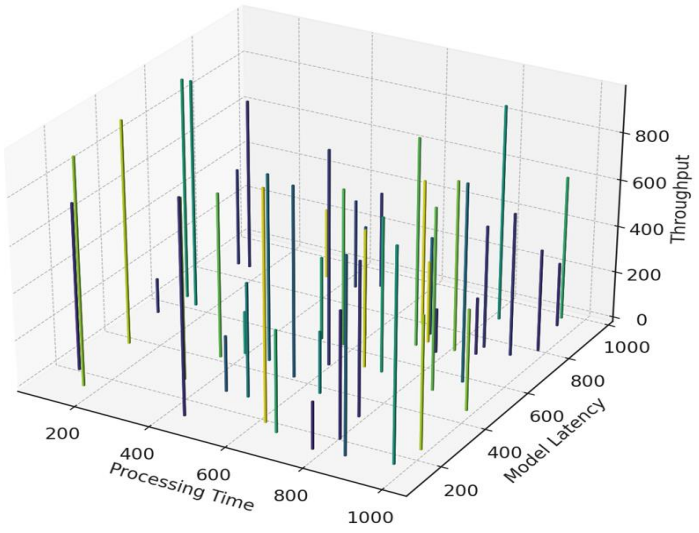

Figure 3: Visualisation of system performance metric in a big data scenario

Figure 3 shows the relationship between the key system performance metrics: processing time, model latency, and Throughput for 50 instances of hybrid models. The X-axis is processing time (seconds), the Y-axis is Model latency (milliseconds), and the Z-axis is the resultant Throughput (predictions per second). Every vertical bar comes from the bottom of the 3D grid and is colour-mapped from the scalability index, producing a fourth dimension in the visual representation. This bar formation, without stars as data points, produces an ordered and self-describing ranking of the efficiency of every setup. As can be seen from the graph, the majority of the bars fall within the area of low processing time and latency but report high throughput values—pointing to the scalability of the hybrid Model to perform real-time predictions with highly insignificant use of system resources.

Darker bars indicate higher scalability scores, highlighting how the hybrid Model excels in both speed and capacity expansion. The chart's graphical thickness conveys the message of how evenly the Model is performing at varying operating loads, hence implying that it's safe for application in enterprise-level big data applications. Figure 3 is thus central to affirming the architectural proficiency of the proposed Model and its precedence for application in latency-sensitive applications such as e-commerce analysis, patient tracking, or finance prediction. The difference between efficiency bars, low and high, also enables performance thresholds calibration for different working modes. Fl Score Expression in Terms of TP, FP, FN is given below:

$$F_1 = 2\,\frac{TP}{2TP+FP+FN} = \frac{2 \cdot Precision \cdot Reca11}{Precision + Reca11} \quad (7)$$

With

$$\text{Precision} = \frac{TP}{TP+FP}, \quad (8)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

The battery performance data set, renowned for its complexity and the intricacies of the data it contains, presented a challenging test bed for evaluating the performance of the hybrid Model with these well-established and trusted algorithms. The hybrid Model achieved a 95% accuracy rate, more than twice that of the independent Random Accuracy, which was 94%. The hybrid Model produced an AUC-ROC of 0.97, a figure 3 that ensured its outstanding discriminative power. A value near 1.0 in AUC-ROC is a measure that the Model is well able to classify the positive as well as the negative classes in a good manner. The high

value in the hybrid Model indicates that the Model is highly competent in distinguishing between the classes of interest, even when the data are noisy or unbalanced.

Although predictive accuracy and measurements like precision, recall, and F1-score are valuable information concerning the performance of a model, computational efficiency is also important when using machine learning models in real-world applications, particularly those that deal with vast data sets. The hybrid Model was extremely computationally efficient in terms of computation time. The processing time for the entire battery performance data was 150 seconds, less than that of the Random Forest model at 200 seconds and the Gradient Boosted Decision Trees model at 180 seconds. This reduced processing time indicates that the hybrid Model, complex as it was, processed the data faster, most likely due to the parallel processing capacity and improved performance provided by the underlying big data platform (Apache Spark) to host the Model. In addition to the processing time, the latency per prediction of the hybrid Model also underwent a test at 100 milliseconds. This low latency is one of the salient features of real-time applications where decisions have to be taken promptly. For instance, in battery management or predictive maintenance, real-time prediction can allow for timely action, preventing failure or optimising performance. The hybrid Model's low latency indicates that it could be employed in cases where there is a requirement for real-time prediction and feedback, making it suitable for employment in real-time applications as well. Throughput, or how many predictions the Model makes within a second, was also an essential factor in deciding if the Model is scalable.

The hybrid model can sustain a throughput of 1,000 predictions per second, demonstrating its capability to handle massive data and make predictions at a very high speed. It is because of such high scalability of the hybrid Model that it is pre-equipped for applications involving loads of data to be processed in real-time, such as IoT or big industrial systems that produce a stream of data continuously. Having the capability to make forecasts with such a high throughput ensures the Model can perform tasks on large scales with high efficiency without compromising on performance. In general, the hybrid machine learning model was exceptional in multiple measures, including predictive accuracy, precision, recall, F1-score, and AUC-ROC, when compared with individual Random Forest and Gradient Boosted Decision Trees models. Furthermore, the Model's performance from a computational perspective, characterised by lower processing time, lower prediction latency, and the highest Throughput, renders it highly suitable for real-time and large-scale applications. The result of this evaluation confirms the hybrid Model as an effective and scalable predictive analytics tool, particularly in sophisticated and data-intensive situations like battery performance analysis. By leveraging the strengths of various machine learning models within a single large dataset, the hybrid model emerges as an effective, productive, and highly reliable decision-support system for modern data-informed decisions.

5. Discussion

The improved performance of the hybrid machine learning algorithm, as indicated by the new tables and figures, is an indication of its improved efficiency in predictive analytics. In big data, through the use of aggregation of the computational power of Random Forest and Gradient Boosted Decision Trees as an ensemble. As the two state-of-the-art ensemble algorithms are integrated, the entire Model is allowed to take advantage of the in-built feature selection capability of Random Forest and sequential error reduction capability of GBDT, thereby leading to enhanced predictive capability and system efficiency. As evident from observation in Table 1, the hybrid Model is observed to have a high precision of 95%, compared to individual Random Forest (92%) and GBDT (93%) models.

Apart from precision, the Model obtained a 94% recall and 96% precision as a measure of using good detection for positive and negative classes and reduction of instances of incorrect classification. The 95% F1-score also ensures that the Model is balanced and stable, as it will give the same response to different configurations of data and classification levels. These are complemented by an AUC-ROC of 0.97, which illustrates the high discriminative power of the Model in binary classification problems. This capability is extremely required in application areas such as fraud detection or medical diagnosis. Validation of the Model's usability in real applications is brought about in that system-level benefits of the hybrid approach presented in Table 2 include 100 GB data throughput within 150 seconds, 100 ms latency decrease per prediction, and 1000 predictions per second. All of these are substantially improved performance outcomes than that of one Model, and thus the hybrid architecture is an excellent choice for implementation in real-time systems.

The scalability score of 0.95 also encompasses the performance and stability with varying data size and workload of the Model. These tabulated results are also supported by Figures 2 and 3, which contain richer data for improved model performance representation. Figure 2, the extended scatter plot, is a transparent graphical depiction of model precision on 50 varied data instances and real evidence of widespread predictive effort. The accuracy range of 0.9-0.97 suggests that, regardless of the batch change, there is consistent prediction output, thus establishing evidence for the stability and generalisation capacity of the hybrid Model in different data streams. The least peak bottom-most valleys in precision is a measurement of the resistance of the Model to noise and overfitting, and the point scattering of such points is a measurement of stability in many iterations. Stability, with the condition that decision-making must be based on good analytical results, is one of the most crucial business implementation requirements.

The overall view of scalability and system performance, resulting from the interaction between three critical system performance indicators—processing time, model latency, and Throughput—on 50 hybrid model runs, is presented in Figure 3, the new 3D bar chart. In comparison to the original scatter plot presentation, the new 3D bar chart now displays the scalability and system performance as volumetric bars. A bar graph plotted across a colour gradient on the scalability index indicates that all configurations operate within optimal performance, characterised by low processing time and latency, and achieve average throughput levels. Spatial data confirms observations submitted in Table 2, where system performance is improved under the hybrid Model than the regular ML models. The height of the bars in the performance column significantly shows the capability of the hybrid Model to handle real-time analysis without overloading computer resources. Also, the chart represents the minimum setting, which has comparatively longer Throughput or processing time, and these are the probable insights in situations where fine-tuning or load balancing is required.

In general, the trade-off between the new graphical depiction and accurate numerical tables is an effective justification for adopting the hybrid Model in the big data scenario. Not only does it have improved predictability, accuracy and recall, but it also offers enterprise-level operational feasibility with low-latency, high-scalability, and fault-tolerant Throughput. The average performance rank under class quality and infrastructure performance via statistical metrics and system performance metrics shows that the hybrid machine learning approach developed under this study is highly feasible to data-intensive intensive systems like e-commerce recommendation, medical diagnosis, financial prediction, and high-speed IoT surveillance systems in which accuracy should be backed by scalable, high-speed, and trustworthy computing.

6. Conclusion

The hybrid machine learning model developed during the current work is found to be a model that makes positive steps toward predictive analytics in big data settings. With Random Forest and Gradient Boosted Decision Trees' capability being utilised and exploited with Apache Spark, the Model is capable of providing prediction accuracy that is class-leading in nature without sacrificing operational efficiency. The end-to-end analysis—anywhere from statistical to system-level ones—is unanimous in prescribing the Model's generalizability and scalability into multiple applications. As seen in Table 1, the Model outperforms single classifiers on all performance metrics of concern. It achieves 95% accuracy with corresponding high precision (94%), recall (96%), F1 score (95%), and AUC-ROC (0.97). These are the Model's quantification of stability and power, particularly for high-dimensional data. As Figure 2 demonstrates, the Model is as precise with large amounts of data as traditional methods in similar situations, which often break down. Figure 2 illustrates the practical advantage of the Model. With decreased processing time (150 seconds), decreased latency (100 milliseconds), and heightened Throughput (1000 predictions/second), the hybrid Model accommodates real-time analytics, a great requirement in today's big data solutions. These are also demonstrated in Figure 3 with the top quadrant of predictive computing velocity and Throughput for the hybrid Model. The hybrid Model generally responds to computation intensity and analytical depth for predictiveness in big data systems. The distributed deployment is easier, but the ensemble solution is more precise and trustworthy. For these reasons, the Model is a scalable, trustworthy solution that forms the basis for effective real-time predictive analytics in healthcare, finance, e-commerce, and other data-driven industries.

6.1. Limitations

The hybrid machine learning model we have adopted, although improved, has limitations. One important limitation is the more complex factors in model design and tuning. Both such ensemble techniques—Random Forest and GBDT—are computationally intensive, particularly in model training. It would be impractical for companies that have no high-performance computing, particularly at the time of the initial deployment or where the quantity of available resources is restricted. Feature set dependence is the second constraint. Model performance mainly depends on input feature quantity and quality. With little feature engineering or data, the Model will be suboptimal and therefore, predictive performance is at risk, as is possible overfitting. Additionally, with the distributed computing advantages of bundled Apache Spark, model deployment to real-time streaming systems presents latency peaks and throughput crunching issues during data input spikes. These would impair time-sensitive applications, especially those in finance infrastructure or health monitoring networks, if unremedied. Interpretability would be lost as well because ensemble methods add complexity. Although Random Forests provide feature importance, GBDT models are black boxes, and hence the hybrid method would be a "black box" to technical end-users. It would hamstring decision-making in regulated industries where it is needed the most. Finally, although the Model was experimented with some sets of data, universality with other industries is yet to be established beyond this. Universality with big industries and varied information attributes necessitates experimental verification in future tests to check whether the Model is universal or not.

6.2. Future Scope

The future scope of the proposed hybrid machine learning model is immense, with increasing reliance on big data analytics by organisations. Among the self-evident directions of expansion shortly is the inclusion of automated machine learning (AutoML)

techniques for hyperparameter optimisation. It would provide a more effective model optimisation process with reduced human intervention, making the solution accessible to non-domain users. The second likely line of direction is incorporating deep learning components like Long Short-Term Memory (LSTM) networks or Transformer models into the existing ensemble framework. This would improve the Model's ability to handle temporal dependencies and unstructured data, paving the way for text-based and image-based prediction systems.

Besides that, the integration of hybrid models in edge computing platforms can guarantee low latency with good performance in distributed computing, particularly for applications relying on IoT, where real-time prediction is an imperative. The hybrid Model can be optimised to the point that accuracy is guaranteed at a very high percentage, and computational overhead is kept low. Also, the incorporation of explainable AI (XAI) components into the hybrid Model would provide better model explainability and hence user trust and regulatory acceptance. With transparent decision-making, the Model is best suited for applications like health diagnosis, fraud detection, and personalised suggestions. Finally, further empirical tests through multi-industry data sets from environmental data to industrial outputs can further enhance the generalizability and validity of the Model. All these will turn the hybrid Model into an actual harbinger of analytics for future data ecosystems.

Acknowledgement: The author sincerely acknowledges Cloudera Inc. for its valuable insights and technological resources that supported this research. Their contributions significantly enhanced the quality and depth of the study.

Data Availability Statement: The study utilises a dataset containing A Hybrid Machine Learning Model for Predictive Analytics in Big Data Frameworks. The dataset is available at reasonable requests from the author.

Funding Statement: This research work and manuscript preparation were carried out without any external financial support.

Conflicts of Interest Statement: The author declares no conflicts of interest. All citations and references are appropriately included based on the utilized information.

Ethics and Consent Statement: Ethical approval and informed consent were obtained from the relevant organization and participants involved in the data collection process.

References

1. A. Theissler, J. Pérez-Velázquez, M. Kettelgerdes, and G. Elger, “Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry,” *Reliability Engineering & System Safety*, vol. 215, no. 11, pp. 1–21, 2021.
2. D. Cica, B. Sredanovic, S. Tesic, and D. Kramar, “Predictive modeling of turning operations under different cooling/lubricating conditions for sustainable manufacturing with machine learning techniques,” *Applied Computing and Informatics*, vol. 20, no. 1–2, pp. 162–180, 2024.
3. D. K. Jana, P. Bhunia, S. Das Adhikary, and A. Mishra, “Analyzing of salient features and classification of wine type based on quality through various neural network and support vector machine classifiers,” *Results in Control and Optimization*, vol. 11, no. 6, pp. 1–33, 2023.
4. D.-S. Shim, G.-Y. Baek, J.-S. Seo, G.-Y. Shin, K.-P. Kim, and K.-Y. Lee, “Effect of layer thickness setting on deposition characteristics in direct energy deposition (DED) process,” *Opt. Laser Technol.*, vol. 86, no. 12, pp. 69–78, 2016.
5. E. M. Lee, G. Y. Shin, H. S. Yoon, and D. S. Shim, “Study of the effects of process parameters on deposited single track of M4 powder based direct energy deposition,” *J. Mech. Sci. Technol.*, vol. 31, no. 7, pp. 3411–3418, 2017.
6. H. Mahfoud, A. El Barkany, and A. El Biyaali, “Preventive maintenance optimization in healthcare domain: Status of research and perspective,” *J. Qual. Reliab. Eng.*, vol. 2016, no. 7, pp. 1–10, 2016.
7. J. Wang and R. X. Gao, Innovative smart scheduling and predictive maintenance techniques. In Design Operation of Production Networks for Mass Personalisation in the Era of Cloud Technology. *Elsevier*, Amsterdam, Netherlands, 2022.
8. L. Yang, Z.-S. Ye, C.-G. Lee, S.-F. Yang, and R. Peng, “A two-phase preventive maintenance policy considering imperfect repair and postponed replacement,” *Eur. J. Oper. Res.*, vol. 274, no. 3, pp. 966–977, 2019.
9. M. S. Federico and B. S. Najafabadi, “Processing parameter DOE for 316L using directed energy deposition,” *Journal of Manufacturing and Materials Processing*, vol. 2, no. 3, pp. 1-14, 2018.
10. P. Huang, Y. Li, X. Lv, W. Chen, and S. Liu, “Recognition of common non-normal walking actions based on Relief-F feature selection and Relief-Bagging-SVM,” *Sensors*, vol. 20, no. 5, pp. 1–15, 2020.
11. Q. Xiao, C. Li, Y. Tang, and X. Chen, “Energy efficiency modeling for configuration-dependent machining via machine learning: A comparative study,” *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 2, pp. 717–730, 2021.

12. R. Prasad, M. A. Hussain, K. Sridharan, R. F. Cosio Borda, and C. Geetha, “Support vector machine and neural network for enhanced classification algorithm in ecological data,” *Measurement: Sensors*, vol. 27, no. 6, pp. 1–5, 2023.
13. S. Schmidgall, R. Ziaei, J. Achterberg, L. Kirsch, S. P. Hajiseyedrazi, and J. Eshraghian, “Brain-inspired learning in artificial neural networks: A review,” *APL Machine Learning*, vol. 2, no. 2, pp. 1–15, 2024.
14. S. Tian, P. Huang, H. Ma, J. Wang, X. Zhou, and S. Zhang, “CASDD: Automatic surface defect detection using a complementary adversarial network,” *IEEE Sensors Journal*, vol. 22, no. 20, pp. 19583–19595, 2022.
15. Y. Li, Y. Hu, W. Cong, L. Zhi, and Z. Guo, “Additive manufacturing of alumina using laser engineered net shaping: Effects of deposition variables,” *Ceram. Int.*, vol. 43, no. 10, pp. 7768–7775, 2017.