Review Article

Web-Centric Cloud Framework for Real-Time Monitoring and Risk Prediction in Clinical Trials Using Machine Learning

Ravi Shankar Garapati *[ID]

Frontend Engineer, Intone Networks Inc, Iselin, NJ 08830, USA

*Correspondence: Ravi Shankar Garapati (yapargam@gmail.com)

Abstract: Advances in web-centric cloud computing have facilitated the establishment of an integrated cloud environment connecting a wide variety of clinical trial stakeholders. A web-centric cloud framework is proposed for real-time monitoring and risk prediction during clinical trials. The framework focuses on identifying relevant datasets, developing a data-management interface, and implementing machine-learning algorithms for data analysis. Detailed descriptions of the data-management interface and the machine-learning processes are provided, targeting active clinical trials with therapeutic uses in cancer. Demonstrations utilize publicly available clinical-trial data from the ClinicalTrials.gov repository. The real-time monitoring and risk prediction systems were assessed by developing five supervised-classification-machine-learning models for trial-status prediction and six unsupervised models for patient-safety-profile assessment, each representing a different phase of the clinical-trial process. All supervised models yielded high accuracy and area-under-the-curve values at the testing stage, while the unsupervised models demonstrated practical applicability. The results underscore the advantages of using the trial-status algorithm, the patient-safety-profile model, and the proposed framework for performing real-time monitoring and risk prediction of clinical trials.

1. Introduction

Real-time monitoring and risk prediction can improve the management of clinical trials. A web-centric cloud framework for real-time clinical trial monitoring and risk prediction using machine learning is designed and implemented. Machine learning has become an invaluable tool in many disciplines, including healthcare. Data sources such as administrative databases, electronic health records, and real-time data from wearable health devices are being exploited to address a wide spectrum of clinical needs. New algorithms are being developed to analyze growing data sets in an effort to create better models that approach human cognition. With the extensive adoption of such algorithms, real-time clinical trial monitoring has emerged as crucial in a clinical trial's data management. Origins of these algorithms can be traced to many sources, including neural computation—one of the oldest and most intriguing sources that aspire to simulate the problem-solving abilities of humans. Over the years, the Artificial Neural Network has received widespread popularity due to its robust self-learning capabilities that rob it of

the need for complex explicit programming. Its performance mimics the behavior of biological neurons in the human brain. Consequently, the range of humanlike tasks that can be performed by a well-trained network are almost endless, supporting predictions, dermatology, image analysis, radiology, and procedural medicine, all of which require various levels of expertise for accurate and efficient interpretation and decision-making.

1.1. Objectives and Scope of the Study

Real-time monitoring and risk prediction have become important in the context of clinical trials. Machine learning plays a central role in the emerging era, especially in health care. Various machine learning algorithms can be applied to the different phases of clinical trials since a large amount of clinical trial data is generated and stored in health data warehouses. The higher patient-risk factors can be identified in earlier phases by applying machine learning algorithms on the clinical trial data. Such risk prediction can improve the quality, safety, and efficacy during the clinical trial execution in health-care organizations. The rapidly evolving discipline of machine learning, combined with adequate data, can provide a sophisticated real-time monitoring framework in clinical trials. Data isolation, lack of seamless integration of data sources and applications in real time, lack of security while moving data, lack of advanced monitoring, risk-prediction framework, and high maintenance render these functions inefficient. Cloud computing is emerging as an advanced computing paradigm that offers numerous services over the Internet from a shared pool of configurable computing resources. By combining these emerging technologies, it is possible to design and implement a web-centric cloud framework for real-time monitoring and risk prediction in clinical trials using machine learning. The framework supports web-centric clinical trial risk-prediction services during clinical trial execution.

Figure 1. Clinical Trial Real-time Monitoring and Risk Prediction Framework

2. Background

Clinical trials play a significant role in testing the safety and effectiveness of medical devices, pharmaceutical drugs, and therapeutic treatments. The clinical trial life cycle comprises the successive completion of many stages where the patient progresses. As the medical device pandemic study case demonstrates, real-time monitoring and risk prediction of patients are pivotal components of the monitoring phase, enhancing the life cycle and quality of the clinical trial. Specifically, real-time monitoring aims to promptly detect unforeseen side effects and risks that may have a detrimental impact on patients. The overarching goal of the clinical trial life cycle is to ensure patients' well-being

while concurrently enhancing device performance. Within medical applications, machine learning represents one of the most studied and implemented forms of Artificial Intelligence. Machine learning algorithms facilitate data-driven learning, extraction, classification, clustering, and data-centric risk predictions. Achieving this requires big data collection and appropriate input selection, which directly influences the decision-making accuracy of the algorithms. A web-centric cloud framework is presented here to enable real-time monitoring and risk prediction during the clinical trial life cycle, utilizing smart medical hardware for more accurate data assessment. By assessing potential risks during the clinical trial, patient safety and data accuracy are amplified.

2.1. Overview of Clinical Trials

The role of clinical trials in the ongoing quest for possible cures for numerous diseases ranges from testing and evaluating the efficacy of a candidate drug in an optimal and controlled setting (randomised controlled trials) to determining the pharmacovigilance measure from an adverse event standpoint via observational studies. Real-time monitoring of clinical trial data before, during, and after the trial is extremely important to detect inconsistencies, ensure quality and timely completion, and implement early interventions to prevent harmful or untoward events—especially for randomised controlled trials. Traditionally, this monitoring, whether by sponsors or by regulatory authorities, is done at a scheduled time and not on an ongoing basis. Another important role of data collected in clinical trials throughout these different phases is to predict any anomalies or risks associated with the trial that might lead to adverse events for subjects/patients involved, a high dropout rate, or even loss of efficacy. Real-time monitoring of clinical trials can be done with a well-organised data lake or warehouse, and anomalies in the data can be detected with the help of machine-learning algorithms. A web-centric software framework is proposed that offers a cloud-based real-time monitoring dashboard and an integrated module with machine-learning algorithms for the early detection and prediction of risks in clinical trials. An end-to-end data-simulation module is incorporated to cover the entire life cycle of clinical trials. Data sourced from different simulated entities assumed to be located in different parts of the world is brought onto a central cloud storage facility and displayed live on the dashboard for seamless monitoring by all concerned parties involved. The risk prediction module is implemented with different machine-learning algorithms—both supervised and unsupervised—applied on publicly available clinical-trial data sets and separate simulated data sets. These algorithms are designed to calculate the risk associated with the principal investigator, the high dropout of patients, or the loss of efficacy, and the model with the best performance is selected and integrated into the facilities management module for continuous risk assessment. The results of the framework have demonstrated how risks in clinical trials can be predicted upfront and appropriate preventive steps implemented early enough to lessen any adverse impact.

2.2. Importance of Real-Time Monitoring

Trials are inherently risky ventures. Participant enrolment may be delayed. Trial sites can be slow to start. Adequacy of the recruited sample may be delayed. Adverse events may occur that require additional monitoring and/or changes to the protocol. It is conventional to monitor and manage risks on a quarterly or monthly basis, depending on the risks and also the phase of the trial. The impact of a lapse in recruitment is dependent on other sites picking up the loss of recruitment. The impact of an adverse event on the trial may be minimal or it may be significant enough to require variation to the conduct of the trial for all future participants. Traditionally, these   trials were monitored using predefined patient-reported sheets. Nowadays, everything is digitized and electronic data are readily available. These data can be further utilized effectively to monitor clinical trials in real time regardless of the source of the data. These sources can

be from real-time clinical databases or from wearable devices that are worn by the patients. These devices can monitor the ECG, blood oxygen level, blood sugar level, etc., and can automatically prepare the respective reports and can be consumed by the back-end system.

Equation 01: Data stream rate (per device type i):

$$Ri = N_i f_i s_i (1-d_i) c_i \Rightarrow R_{total} = \sum_i R_i, \tag{1}$$

where N_i = device count, f_i = sample rate (Hz), s_i = payload (bytes), d_i = drop rate, c_i = compression ratio.
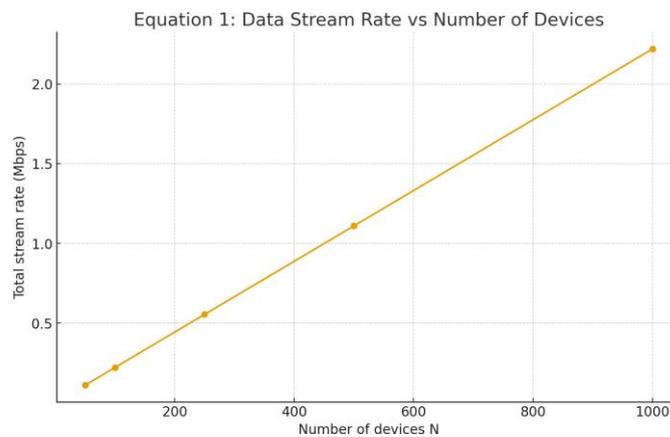

Figure 2. Data Stream Rate vs Number of Devices

N devices	Total stream rate MBps	Total stream rate Mbps
50	0.013876	0.111008
100	0.027752	0.222016
250	0.06938	0.55504
500	0.13876	1.11008
1000	0.27752	2.22016

2.3. Machine Learning in Healthcare

Healthcare is broadly defined as any intervention that aims to improve health. Medical care is the diagnosis, treatment, and prevention of disease, illness, injury, and other physical and mental impairments in people. Delivering the right care requires a great deal of data analysis. Data comes from a variety of sources including wearable sensors, medical images and documents, FDA databases, as well as EHRs and other medical records. Intervention predictions can be made through the application of supervised learning techniques   such as classification, regression, and Apache Spark, along with unsupervised learning techniques including clustering and autoencoders. By utilizing these machine learning and deep learning models, it becomes possible to identify patients who are susceptible to specific diseases, people who are less likely to use hospital services, or individuals who are likely to be re-admitted. This method can also predict the intervention needed for specific patient groups or diseases, estimate the success probability of different therapies, forecast the potential outbreak of diseases in specific regions, and delineate probable disease regions on the planet map. These predictions can

be used to provide the right intervention at the right time for the right set of patients, thus optimizing resource utilization and reducing healthcare costs.

3. Framework Design

A web-centric cloud framework for real-time monitoring and risk prediction in clinical trials at a large healthcare provider is proposed. Clinical trial monitoring is viewed as a machine-learning supervised classification problem. The framework can support various supervised and unsupervised classification techniques. Several risk-prediction models are developed to reduce risk in clinical trial monitoring and failure. Clinical trials form the backbone of the healthcare sector. In clinical research, medical interventions are tested and proven to do more good than harm. With the application of advanced techniques used in other domains, such as for real-time data acquisition and predictive analytics, clinical trials in healthcare can provide better oversight to the process and outcome.

3.1. Architecture Overview

Clinical trials play an indispensable role in healthcare research. Real-time clinical trial monitoring not only ensures human participant safety but also provides high-quality and timely observational data for statistical analysis. Continuous progress in machine learning has made risk prediction increasingly more feasible in the environment of smart healthcare, thereby reducing risk factors. Section 3.1 presents an overview of a framework that leverages these technologies to facilitate real-time monitoring and risk prediction for clinical trials. Clinical trial progress data are gathered from a variety of sources, including progress reports, patient records, laboratory documents, clinical trial protocols, investigator brochures, serious adverse events, and other related tracking documents. Such documents often exist as unstructured text, images, and other unorganized digital formats. The framework relies on cloud computing to offer on-demand computational and storage resources; a cloud-based Web application supports real-time monitoring and risk prediction of clinical trial progress. Machine learning assists in pinpointing risk factors and understanding their association with clinical trial progress.

3.2. Cloud Infrastructure

A cloud-centric infrastructure, supported by Amazon Web Services (AWS), enables the integration of data from multiple heterogeneous sources. With more than 200 fully featured services accessible from data centers worldwide, AWS offers remarkable flexibility and incremental monthly billing. Employing Amazon Simple Storage Service (Amazon S3) as the cloud-storage system underscores storage efficiency and exceptional scalability. Various health information sources—such as ClinicalTrials.gov, PubMed, World Health Organization (WHO), Food and Drug Administration (FDA), and Centers for Disease Control and Prevention (CDC)—contribute to   risk prediction and real-time monitoring. Amazon Athena facilitates ad hoc query and analysis of real-time monitoring data sets stored in Amazon S3. Using standard SQL syntax, healthcare specialists execute queries directly on these data sets within Amazon S3, obtaining results in seconds without complex combination or movement of data. This approach eliminates the need for loading into specialized analytical systems. Operational monitoring results store within an Amazon RDS (Relational Database Service) PostgreSQL database and run daily risk-prediction models. PostgreSQL database snapshots reside in AWS Zone S3 and are protected from single data-center failures. Figures illustrate the risk prediction and operational monitoring processes, encompassing ingestion, data preparation, model building, execution, and performance evaluation.

3.3. Web-Centric Components

The previous three subsections explored the data sources, machine learning algorithms, and required cloud infrastructure. The current section describes the web-centric framework components that support the seamless flow of data from the sources to the cloud infrastructure, machine learning algorithms, and result visualization. An ecological model is often used in population health studies, incorporating multiple levels of influence that affect individual and community health, namely, a combination of individual, relationships, community, and societal factors. Similarly, the web-centric framework supports data flow from individual to societal levels of information. At the individual level, smart devices are rapidly emerging in the area of the Internet of Health Things, allowing individuals to enhance their health and communicate information to others. Clinical trial participants, research teams, and sponsors benefit from real-time monitoring of healthcare studies via an ecosystem of interconnected devices and applications capable of sharing health data. Real-time monitoring reduces or eliminates types of human error such as distant batch review, non-uniformity of patient record check, visual error, and selection error. At the relationship level, Monkeypox clinical trial participants are in close contact with each other. Participants' social circles, such as family, friends, co-workers, or neighbors, also play a significant role in participants' decisions to enroll. The strongest effect of social support on health is that it makes it easier to cope with stress and prevents isolation, depression, and withdrawal. At the community level, the risks associated with clinical trials must be addressed based on risks at community level, including its housing, socioeconomic status, crime, culture and language, transportation, and availability of day care services [1].

4. Data Collection and Management

Data from virtually all application domains are increasingly becoming available in digital form. Much of this data exists in the form of tables, making the relational data model an attractive vehicle for representation and manipulation. Clinical trials activity generates an enormous amount of data. The type of data generated in clinical trials can be demographics, diagnosis, drug coverage, laboratory test results, prior medications, adverse events, etc. Monitoring of a clinical trial can be performed starting from a synthetic example. For a real-time patient management system, data from a hospital is necessary. For example, a real-time hospital scenario of the inpatient department containing information pertaining to the patient, the diseases from which the patients are suffered, the medical tests performed on the patient, is more suitable [2]. The data is subjected to a simplistic categorization of patient, doctor, and diagnosis to abstract general information about patient, history of diseases, drugs prescribed, and the medical tests undertaken. Such information is usually involved in real clinical trial examinations. Any clinical trial involves confidential data of participants, such as enrollment information, personal identity details, drug history, season of getting illness, etc., which raises serious privacy and data security concerns. Privacy protection of clinical trial data against any kind of leakage has become a fundamental issue in data publishing and sharing.

Equation 02: Expected monitoring latency (micro-batch):
Arrivals uniform within $\Delta \Rightarrow \mathbb{E}[t_{trigger}] = \Delta / 2\Delta \Rightarrow E[t_{trigger}] = \Delta / 2$.

Batch size $n = \Delta \lambda$, capacity $c \Rightarrow \mu$ batch service $n_b = \Delta \lambda Tb = \mu \lambda \Delta Tb$ and average in-batch offset $n_b / 2Tb / 2$.

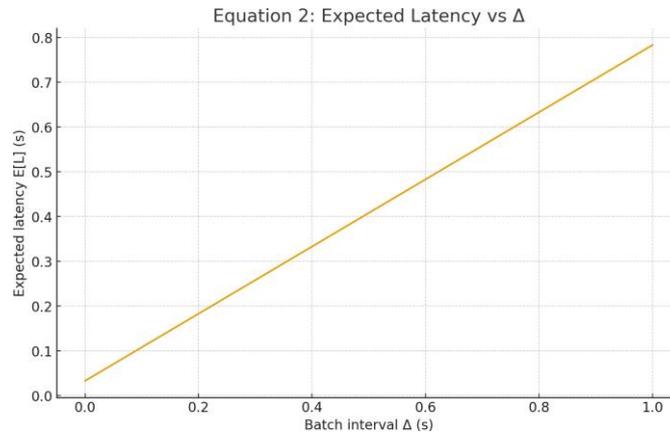$$\mathbb{E}[L] = t_{ingest} + t_{net} + t_{store} + 2\Delta + 2\mu \lambda \Delta \qquad (2)$$

Figure 3. Expected Latency vs Δ

batch interval Delta(s)	expected latency E[L](s)
0.0	0.033
0.2	0.183
0.4	0.333
0.6	0.483
0.8	0.633
1.0	0.783

4.1. Data Sources

Effective real-time monitoring and risk prediction in clinical trials demand seamless and timely access to data from various sources. One primary data source is the electronic Clinical Development Database (eCTD), which stores patient health records. Data originating from wearable sensors, continuously monitoring physical activities such as heart rate, skin temperature, blood glucose, and blood pressure, can also provide critical insights. Such sensor data may either be stored within an eCTD system or integrated in real time with the system for hybrid disease prediction. Data generated using Internet of Things (IoT) devices represent another valuable source. All the data collected from these sources are transmitted to the web-centric cloud framework for processing and risk assessment [3]. Cloud infrastructure and supporting services facilitate the distribution of data to independent Web services that perform various analyses and transformations. The architectural openness of these Web services caters for future enhancements. The framework architecture enables the integration of additional data collection endpoints — whether integrating with Twitter feeds or sensor-based streams — while feeding back future predictions to the data stores. Personal healthcare and fitness information are inherently sensitive. Privacy and security must be front and center during any architecture design process. Access to services must be protected, ideally provoking authentication and authorization processes before any analytical or monitoring services are made available to users.

4.2. Data Integration Techniques

A clinical trial is a prospective research study or investigation that involves human beings as subjects and is performed with their informed consent. It is a very important part of healthcare business operation. Though historically, clinical trials were viewed as a later stage procedure in the overall drug discovery process, trial data are now used not only to help regulatory authorities in establishing safety and efficacy of study agents, but also by pharmaceutical companies to make decisions regarding business operations. Hence it is very important to have real-time monitoring and predictive analysis of clinical

trial data [4]. Machine learning algorithms have been proved to be very successful in various domains including healthcare. Since clinical trial entities have a rich set of data sources, applying machine learning algorithms can help to visualise the actual progress of the clinical trial and predict potential risks. The main objective of this work is to integrate supervised and unsupervised machine learning algorithms in a web-centric cloud framework that caters to real-time monitoring and risk prediction of clinical trials and thereby help clinical trial sponsors to take timely decisions.

4.3. Data Privacy and Security

In spite of the potential benefits, it is important to verify the risk of sensitive data leakage before data migration to the cloud. Accordingly, there were some privacy issues raised. A significant factor that influences whether the business model is considered safe or practical is the ability to protect user privacy. Subsequently, with regard to privacy concerns, access controls, encryption, and authentication protection protocols were proposed. Data privacy can be ensured by maintaining secure and effective two-factor authentication, as well as by using hybrid cloud surveillance enforcing only authenticated users to gain access to cloud data. In addition, a privacy protocol that protects patient confidentiality without relying on a trusted party can be used. Due to both the changing nature of the clinical trials as well as the altering state of patients during a study, test data are examined and updated numerous times. Therefore, it is important to verify not only the patient's clinical data but also the altered data during the alterations and transmission. Data can also be encrypted with attribute-based signcryption, which guarantees confidentiality and removes the identity of the sender or receiver without a trusted third party. Furthermore, a geometric privacy model could be designed to avoid data attacks while uploading and scrutinizing clinical data on the cloud. Data security within the cloud can be provided with effective protection based on the private cloud while uploading any clinical data. Additionally, the use of machine learning models, such as several encrypted deep-learning models, can contribute to a higher degree of security in the cloud [5].

Figure 4. Clinical Trial Management and Risk Prediction Framework

5. Machine Learning Algorithms

Clinical trial data come in many forms (primarily numerical and text, but also images and voice). Machine learning algorithms have the advantage over classical statistical models in that they can be tailored to specific data sets and adjusted in a process of training and testing using actual clinical trial data. Supervised learning techniques can be used to classify clinical trial outcomes and project potential risks, while unsupervised learning techniques can examine the collected data and group elements together to highlight shared characteristics. A key application is the use of supervised learning

techniques to predict clinical trial risk. Three algorithms were considered: kNN, Random Forest, and Logistic Regression [6].

5.1. Supervised Learning Techniques

The objective of risk prediction is to identify risk factors such as lost visits, patient dropout, insufficient patient enrolment, delayed treatment, budget overrun, and low data quality in clinical data. Labelled clinical trial data was collected and several supervised learning techniques were evaluated for risk prediction. In supervised classification, two outputs are possible, for example, the loss of visits can be either "loss" or "no loss" of visit. The availability of risk factors affecting the clinical trials, role of patient in clinical trials, and the conditions of clinical trial play a vital role in the selection of the appropriate model. Hence, several algorithms have been evaluated for their performance in risk prediction. Logistic regression fits a generalized linear model to estimate the probability of an event using the logistic function. Support vector machines (SVM) use hyperplanes to separate data into different classes—such as "risk" or "no risk." The neural network model is a classifier suitable for estimating risk in clinical trials. Perceptrons, which are a single-unit neural network, comprise discrete neurons as building units, transforming input vectors to produce desired outputs. Adaptive boosting (AdaBoost) serves as a meta algorithm to combine several weak classifiers into a stronger one, constructed through combinational training processes.

5.2. Unsupervised Learning Techniques

Unsupervised learning contains no labels and focuses on finding the hidden patterns and connections in the data instances. Unsupervised models include clustering, anomaly detection, and association rule. Clustering is popularly used in exploratory data analysis to find the natural grouping within a set of data [7]. A method is developed to check the quality of healthy volunteers' admission screening and therapist's data entry process. The doctors do not admit volunteers blindly into the clinical trial. There is a huge criteria list to check whether a person is eligible to be admitted in the study or not. For example, if in the trial it is required to have people between 30 years and 45 years, then only people in this age group should be enrolled. If any patient is admitted who is not eligible for this criterion, then it is a child risk in the study. Research suggests that clustering can be used to find the left-out values. K-Means clustering algorithm is used. Patients go through various screening procedures before admission, such as urine analysis, ECG report analysis, cholesterol levels checking, thyroid level checking, SGOT, SGPT analysis, gender and age check, enquiring about disease history, allergy checking, smoking habits, or fall in the category of high risk individuals as per the particular study need. For example, if a patient suffers from weakness in the muscles and believes that he should not be enrolled in the study, still the operator enters the patient details in the system and admits the patient, then it is an induced risk in the study. This clustering method provides aberrant clusters as the output. Some clusters have two to three volunteers, but other groups have a lot of volunteers. These groups are studied and the volunteers in the smaller clusters are found.

Equation 03: Risk probability (logistic regression):

$$p(y = 1 \mid x) = \sigma(z) = \frac{1}{1 + e^{-z}}, \tag{3}$$

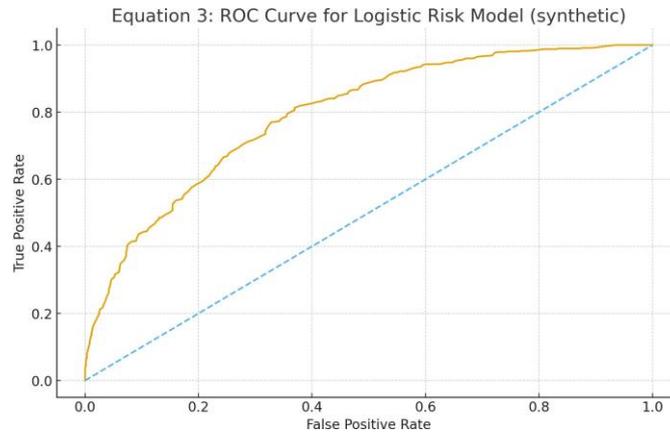$$z = w^{\top} x + b, \; log \frac{p}{1-p} = w^{\top} x + b. \tag{4}$$

Figure 5. ROC Curve for Logistic Risk Model (synthetic)

KPI1	KPI2	p risk	label
1.690525703800356	-1.0873724682195882	0.9425962724666332	1
-0.6254289739667597	-0.6204544998963897	0.4123761367208112	0
-1.5265245318698402	0.0150233356136717	0.125225710332932	0

5.3. Model Selection and Evaluation

Model Selection and Evaluation Near real-time monitoring of risk trends and detection of potential unusual patterns are critical for assessing clinical trial quality. It would be beneficial to develop a risk prediction framework that applies historical risk-based performance data to the historical clinical trial data. The framework employs unbiased feature selection with stability analysis and utilizes supervised and unsupervised machine learning algorithms for risk prediction. The predicted risk scores for key performance indicators are incorporated into a risk-based real-time clinical trial monitoring system and visualized on interactive dashboards [8]. Risk-based Quality Monitoring of ongoing Clinical Trials. Selecting appropriate algorithms and features is essential for achieving stable and accurate risk prediction outcomes. Two distinct clinical trial risk prediction problems are considered: (1) classification of supervised machine learning algorithms for the change status in key performance indicators, and (2) clustering of unsupervised machine learning clustering algorithms to identify clusters of clinical trial sites for the level status in key performance indicators. Several supervised algorithms (logistic regression, linear discriminant analysis, support vector machine, random forest, and XGBoost) and unsupervised algorithms (kmeans, hierarchical, and Gaussian mixture model clustering) are evaluated. For both approaches, all stability scores for feature groups are positive, indicating all features are relevant. The best-performing supervised and unsupervised algorithms for the target key performance indicators—namely median adverse events rating, median cycle time rating, median visit rating, median query rating, and overall risk rating—are determined with regard to the quality classification and level clustering tasks, respectively. Models are assessed with varying training data sizes, and calibrated results demonstrate that the supervised model confidently predicts the risk of unusual patterns.

6. Real-Time Monitoring System

Real-time monitoring ensures patient safety, data reliability, and reporting accuracy, constituting an integral part of good clinical practice. Detecting irregularities and inconsistencies between expected and observed behaviors, often termed as outlier analysis, constitutes a fundamental component of clinical data management. Such anomalies may signal critical clinical events or procedural errors, requiring immediate attention and

potentially prompting action deviations during the trial. The myriad statistical approaches available cater to specific needs within the clinical trial domain [9]. The monitoring system seeks to observe trial progress indicators in near-real time, analyzing these in conjunction with risk indicators to generate alerts when certain risk thresholds are approached. Communication channels are established with trial personnel to endorse these alerts and facilitate corrective measures if required. Performance validation is achieved through a retrospective quality risk assessment of an ongoing clinical trial to which the ML-based monitoring tool is applied. Risk indicators known to correlate with monitoring findings are monitored, and significant associations between risk predictions and the volume of triggered alerts are established. The system harnesses a cloud-based web platform and the monitoring tools described earlier to propose a web-centric cloud framework. The ML models described in the previous section serve as the risk prediction component within the broader monitoring framework [10].

6.1. Monitoring Protocols

Recent advances in cloud computing and machine learning, implemented and deployed using distributed storage systems, offer numerous opportunities in the healthcare domain—particularly in clinical trials. Real-time monitoring and machine learning–based risk prediction can help clinical trial stakeholders perform interventions anytime an adverse event occurs. Effective monitoring of clinical trials for adverse event detection is therefore crucial, and several studies have explored the role of machine learning in risk-prediction analysis systems for detecting adverse events in clinical trials. It is imperative that machine learning methods applied to clinical trials be capable of operating in real time at scale. A web-centric cloud framework is proposed to enable real-time monitoring of clinical trials and risk prediction. Machine learning methods capable of supporting a range of clinical trial–specific applications are also examined. A web-centric cloud infrastructure is designed to support real-time monitoring of ongoing clinical trials and the use of machine learning for risk predictions. The design addresses specific monitoring challenges related to risk prediction of clinical trials, employing commonly used supervised and unsupervised machine learning algorithms, methods for their selection, and performance-evaluation criteria in the context of clinical trials. Machine learning–based risk-prediction models are designed for monitoring that can impose ordinary constraints on clinical trials. Privacy and security are also important aspects of data management in clinical trials; through careful design and implementation, they can be addressed effectively [11].

6.1. Alert Mechanisms

SOP violations may be detected when there is a delay in visit date, an inconsistency in visit dates, or a visit skip. For instance, if the subject is late in appearing for assessment of adverse event (AE) for more than N days, where N is the number of days for tolerable delay in AE visit, an alert will be generated by the web site and sent to the registered email address [12]. The EDC (electronic data capture) is part of the CTMS and protocol deviation risk scenarios can be provided on the EDC as well. If the sites have not submitted the AE data for the subject till the visit is marked as done, the system will generate an alert every day till the AE form is submitted. If the AE form is submitted after the visit is marked as done, the discrepancy number will be raised and the site will be alerted to update the visit status to AE visit done.

Equation 04: Cross-entropy loss (binary) from Bernoulli likelihood $p^{y}(1-p)^{1-y}$:

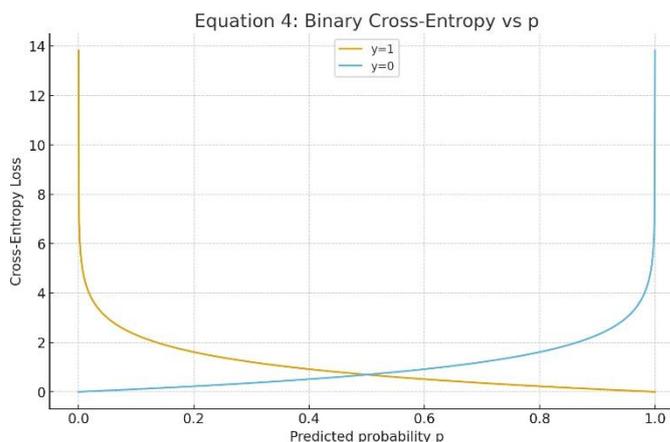$$L = -log\, p y (1-p)1 - y = -[y log p + (1-y)log(1-p)]. \qquad (5)$$

Figure 6. Binary Cross-Entropy vs p

p	Loss when y=1	Loss when y=0
1e-06	13.815510557964274	1.000000500029089e-06
0.0100109899899899	4.6040717904464	0.0100614369151078
0.0190199809809809	3.962265221813817	0.0192031875966616

6.2. User Interface Design

The user interface (UI) development encompasses a web application linked to the Google Cloud Platform (GCP) infrastructure backbone enabling risk prediction and real-time monitoring. User authentication is controlled via Google Cloud Identity and Access Management (Cloud IAM). Designed to be compatible with several Internet browsers, users can visualize both the predicted patient risk and the status of ongoing clinical studies. Figure 6.8 illustrates the main page. End-users can customize risk monitoring and receive alerts for specific clinical trials. Currently, the application supports the monitoring of COVID-19 clinical trials from the ClinicalTrials.gov registry. The documentation of the user interface is available in the online Supplementary Material S4 [13].

7. Risk Prediction Models

Risk prediction has become a highly researched healthcare area. Clinical risk prediction for patients' hypertension, heart disease, and diabetes has been extensively studied using various machine-learning algorithms. Other healthcare domains, including disease risk prediction, risk monitoring, and role of risk in the context of COVID-19, have also been examined. In clinical trials, quality, risk management, and financial management risks play crucial roles in the success of drug product approval and approval cycle times of new product launches—in other words, the focus on patient safety, data quality, and compliance. Clinical trial quality risk was assessed using a hybrid predictive approach that integrated association rules and predictions based on the risks identified during clinical trial vendor selection. Future work could explore automation in the prediction of clinical trial quality, focusing on the moving window concept and length of moving window [14].

7.1. Risk Factors Identification

Clinical trial risk factors have direct effects on the clinical trial risks, and their identification is a critical step in the risk prediction process. A clinically conducted trial generates much trial data from all the three subareas: Clinical Site, Clinical Trial Protocol, and Clinical Trial Subject. Real-time risk estimation of clinical trials requires

comprehensive data, therefore all the above-mentioned three different areas of clinical trials are considered for discussion. However, selection of risk factors can be limited to a clinical site or clinical subject level, if real-time risk estimation of a clinical trial is intended only at either clinical site-level or clinical subject-level, respectively. The identification process relies on a comprehensive trial data set collected from several trial sites for a wide range of clinical studies. Even though the identified risk factors differ depending on the type of cancer under experimentation, a broad range of cancers is considered here for clinical risk prediction in a real-time monitoring environment. The subsequent step of the risk prediction process is the collection of a wide range of data in the respective sub-area. Real-time risk prediction of a clinical trial site requires data related to clinical site risk factors. These factors have also been ranked subjectively through a survey with senior trial members who have years of clinical trial experience. Therefore, a detailed description of the sub-area is given along with the identified risk factors before proceeding to the subsequent section.

7.2. Predictive Analytics

Supervised and unsupervised machine-learning techniques are applied to develop a drug risk-prediction models. Through a risk-assessment analysis of clinical-trial environments, risk factors are extracted and the response of these factors is predicted as a risk-index value using both supervised and unsupervised methods. Prediction accuracy is evaluated with key metrics, including precision, recall, F-measure, accuracy, and area under the curve. A predictive-analytics tool-design approach is presented that applies machine learning techniques to classify ongoing clinical trials, predict performance factors, and generate early warnings during a trial. Multiple classifier models are developed, indicating that logistic-regression and naive-Bayes classifiers provide better performance for clinical-trial classification; naive-Bayes classifiers also yield higher accuracy in predicting identified risk factors [15]. Evaluations suggest that unsupervised clustering methods can simplify high-variance datasets and, in some applications, of supervised-classifier training by grouping similar clinical sites into clusters, thus enhancing classification accuracy. This analysis from the viewpoints of predictive analytics and the risk-management process offers new insights into clinical trials.

Real-time Risk Estimation

Comprehensives data from all three sub-areas
is crucial for accurate risk prediction

Figure 7. Clinical Trial Risk Factor Identification and Estimation

Equation 05: Prediction accuracy & related metrics (for counts TP, FP, TN, FN):

$$Accuracy = NTP + TN, Precision = TP + FPTP, \qquad (6)$$

$$Recall = TP + FNTP, F1 = P + R2PR. \qquad (7)$$
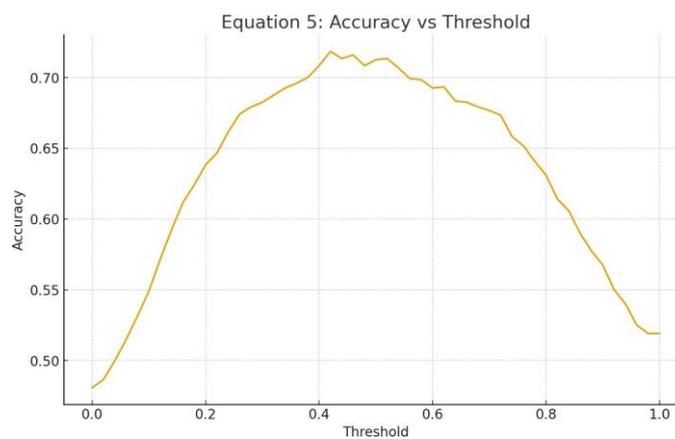
Figure 8. Accuracy vs Threshold

threshold	accuracy	precision	recall	F1
0.0	0.48083333333333333	0.48083333333333333	1.0	0.6494091164884637
0.2	0.6383333333333333	0.5747126436781609	0.9532062391681108	0.7170795306388527
0.4	0.7083333333333334	0.656121045392022	0.8266897746967071	0.7315950920245399
0.6	0.6925	0.7464454976303317	0.5459272097053726	0.6306306306306306
0.8	0.6308333333333334	0.8563829787234043	0.2790294627383015	0.4209150326797384
1.0	0.5191666666666667		0.0	

7.3. Model Validation

Upon the finalization of the clustering algorithm parameters, its suitability is evaluated using an internal validation criterion, specifically the Average Silhouette Width (ASW). ASW measures the quality of the clustering solution by assessing the compactness and separation of the clusters. It examines the cohesion of a particular object in the cluster relative to the separation between the remaining clusters. The measure ranges between "+1" and "-1"; a value close to "+1" indicates that the object is well classified, whereas a value close to "-1" suggests it may be incorrectly assigned to a certain cluster. An ASW value near zero implies that the data point lies on or very close to the decision boundary between two neighboring clusters. The quality of the clusters is evaluated by analyzing the average silhouette width corresponding to the optimal number of clusters. Several actual and placebo clinical trials undergo risk prediction using the K-means clustering algorithm. The preliminary results obtained from the model are further validated by physicians specializing in clinical trial management using real data. A prototype risk prediction software is implemented within a cloud environment to facilitate real-time monitoring of ongoing clinical trials. Such real-time monitoring profoundly reduces the chance of failure by early identification of risk factors and offers protection to the study subjects participating in the clinical trial. The results of this analysis play an important role in improved planning and design of any clinical trial [16].

8. Future Directions

The proposed framework supports clinical trial risk prediction in a real-time monitoring environment. An inspiring direction for the future is to incorporate emerging advances in machine learning techniques. Examples include multitask Bayesian optimization, which has demonstrated notable performance in complex optimization problems with fewer evaluations. Exploring the application of such methods may enhance the efficiency and accuracy of risk prediction in clinical trial monitoring. When integrated into the current cloud infrastructure, these advances have the potential to deliver enhanced capabilities and benefits. As wearable and mobile technologies advance, the scope of clinical trials continues to broaden. Clinical studies naturally expand beyond the confines of dedicated clinical spaces, with participants undergoing interventions in everyday environments. This ongoing shift increasingly blurs the boundary between intervention and nonintervention settings. Future explorations may examine how the proposed monitoring framework can be applied to new and evolving domains in healthcare monitoring, such as smart-home environments for the elderly or for individuals requiring cognitive or physical assistance. In these cases, the quality standards often differ from those governing clinical trials but remain essential for the continued improvement of health and quality of life [17].

8.1. Advancements in Machine Learning

Machine learning—a core branch of artificial intelligence—concerns building and analyzing systems that learn from data. In the healthcare domain, where intelligent decision-making is crucial, machine learning and data mining are applied broadly, addressing problems such as disease prediction, image classification, patient monitoring, and drug toxicity classification. The rapid growth of these fields has generated enthusiasm, recognized potential, and incited enthusiasm among researchers. Machine learning encompasses various types and approaches. Supervised learning employs labeled training data to predict properties of new instances. In contrast, unsupervised learning identifies patterns and regularities within datasets. Both paradigms play vital roles in the context of risk prediction during clinical trials.

8.2. Integration with Wearable Devices

Recent healthcare research increasingly explores artificial intelligence applications and machine learning algorithms that utilize physiological data from wearable technology. Examples include Alzheimer's disease detection, anxiety detection, and COVID-19 infection detection. The application of wearable technology in clinical trials has garnered corresponding attention, as it allows for the real-time collection of physiological data. The availability of such data offers significant advantages for clinical trial monitoring and risk prediction in the pharmaceutical industry. These data can be directly integrated into the web-centric cloud framework to address real-time monitoring and risk prediction requirements [18]. The presented framework capitalizes on the integration of Machine Learning (ML) techniques for the real-time monitoring and risk prediction of clinical trial data. Monitoring functions in clinical trials help prevent potential problems. Unsupervised ML algorithms assess the performance of clinical trial sites and identify other potential issues, subsequently issuing alerts for the appropriate personnel. Core elements of a Risk-Based Monitoring approach focus on risk differentiation and tracking. Identifying and prioritizing services, associated risks, and risk indicators for clinical trial processes enables the development of a risk-based perspective on clinical trial processes and their interdependencies. Key Performance Indicators (KPIs) and Key Risk Indicators (KRIs) monitor performance and risks continuously, with breaches triggering notifications to the relevant business units.

Equation 06: Cloud resource utilization (per resource r):

$$Ur = Cr / Dr, where\ Dr = demand\left(units / s\right), \tag{8}$$

$$Cr = capacity\left(units / s\right). \tag{9}$$
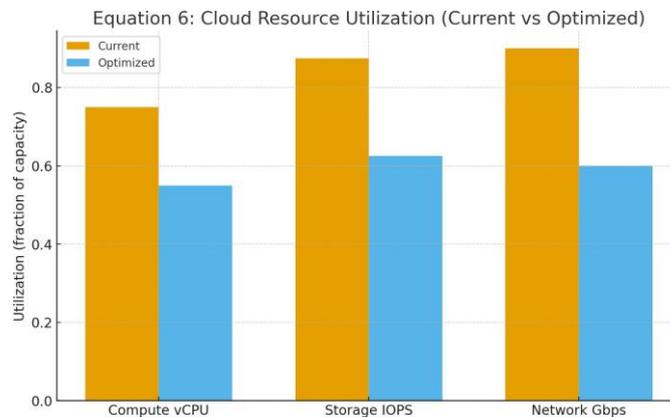

Figure 9. Cloud Resource Utilization (Current vs Optimized)

Resource	Capacity	Current Demand	Optimized Demand	Current Utilization	Optimized Utilization
Compute vCPU	2000.0	1500.0	1100.0	0.75	0.55
Storage IOPS	80000.0	70000.0	50000.0	0.875	0.625
Network Gbps	20.0	18.0	12.0	0.9	0.6

8.3. Expanding to Other Healthcare Domains

As machine learning matures and new algorithms appear, the framework remains responsive to change, with a built-in interface that allows insertion of new, state-of-the-art machine learning implementations. New data sources, such as wearable devices monitoring patient activity and mood patterns, extend applicability beyond clinical-trial

management to Real-World Data analysis. Real-time systems monitor data as soon as they are collected, improving the quality of care in domains such as heart problems and pain management. These analyses are equally important for other conditions—patient comfort must be adapted to the patient's needs in real time. Expanding into areas such as patient rehabilitation and other domains and medical problems can be swift and straightforward. This is facilitated by the underlying infrastructure and the pluggable machine-learning interface, which enable timely and efficient consideration of monitored data at both aggregate and patient-specific levels. Much of this capability is tested by the requirements of Real-World Data, which allow for additional steps and assessments as patient awareness improves. The maturity of machine-learning technologies and the growth of sensor-based medical monitoring devices have shaped the development of an architecture supporting real-time engagement with clinical-trial data. This design ensures consistent data flow and responsive analysis, facilitating timely decision-making and effective risk mitigation [19].

9. Conclusion

The research presented a cloud-centric framework that orchestrates Internet-of-Things, World Wide Web, and cloud computing technologies in a composer, binder, and broker model, and machine-learning techniques in a real-time monitoring and risk-prediction model for clinical trials. The machine-learning component incorporates both supervised and unsupervised-learning approaches. The Grace Healthcare system implemented on Amazon Web Services exemplifies the composer, broker, and binder model, while the monitoring and-prediction system realized the clinical-trial risk-prediction model. It is shown that the framework is capable of monitoring and analyzing data generated by distributed clinical trials in an Amazon Web Services Real-Time Clinical Trial Monitoring System, and that the proposed machine-learning model can identify potential risks and alert stakeholders accordingly. Grace Healthcare represents a cloud-based framework for clinical trials that incorporates both medical IoT applications and trial data analytics. It permits hospital and non-hospital sites to monitor trials in real time by integrating data produced by IoT medical devices. The system can perform patient-adherence analyses and risk predictions using supervised machine-learning algorithms with historical data. The machine-learning model can also identify data patterns and relationships with clinical-trial patient groups using unsupervised approaches. The framework supports distributed clinical-trial investigations, empowers hospitals in trial administration, evaluation, and supervision, and improves product safety and quality by controlling and minimizing patient-risk potentials in real-time clinical investigation scenarios. A cloud centric framework was proposed for clinical-trial monitoring and risk prediction based on the composer, binder, and broker interoperability model. The model leverages Internet of Things, World Wide Web, and cloud-computing technologies in response to the growing complexity and resource requirements of clinical trials. Distinctively, a Google-CloudPlatform–based real-time clinical-trial data-collection mechanism was developed for monitoring distributed investigatory sites. Data records from the real-time mechanism are committed to Amazon-Relational-Databases and are also consolidated with historical data. Within its risk-prediction component, the model integrates both supervised and unsupervised-learning approaches that facilitate patient-profile clustering according to risk level, analyze patient-treatment adherence, predict potential risks, and suggest corrective actions. The Amazon-WebServices–hosted Grace Healthcare demonstration confirms the model's capacity to provide an integrated real-time view of engaged groups/products/patients to clinical-trial officers and risk-analysts at various stakeholder levels [20].
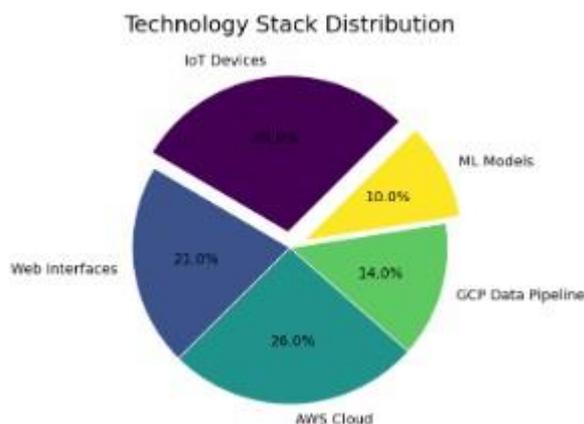
Figure 10. Technology Stack Distribution

9.1. Key Takeaways and Final Thoughts

Real-time monitoring and risk prediction have long been recognized as critical aspects of managing clinical trials. The emergence of advanced machine learning techniques has recently added a new dimension to the practice, enabling the extraction of risks and providing valuable information for decision-makers. Clinicians are capable of assessing potential threats and proactively preventing or mitigating their effects. As machine learning evolves, it fosters the development of a web-centric cloud framework that supports these aspirations. Any clinical trial monitoring system should also adopt a web-centric cloud-enabled architecture to facilitate the smooth flow of relevant information, ensuring that decision-makers always have the necessary data at hand. By integrating these components, a holistic approach emerges, offering a faster, more efficient, reliable, cost-effective, secure, and accountable solution. Expanding the analysis beyond clinical trials, risk factors associated with diseases can be identified to predict potential threats for patients and suggest preventive measures. Recent advancements in wearable devices offer abundant physiological data, which, when harnessed, enable the design of real-time monitoring and risk prediction systems for various healthcare domains.

References

[1] Inala, R. (2022). Engineering Data Products for Investment Analytics: The Role of Product Master Data and Scalable Big Data Solutions. International Journal of Scientific Research and Modern Technology, 155–171. https://doi.org/10.38124/ijsrmt.v1i12.636

[2] Agrafiotis, D. K., Lobanov, V. S., Farnum, M. A., et al. (2018). Risk-based monitoring of clinical trials: An integrative approach. *Clinical Therapeutics, 40*(7), 1204–1212. https://doi.org/10.1016/j.clinthera.2018.06.012

[3] Goutham Kumar Sheelam, "Semiconductor Innovation for Edge AI: Enabling Ultra-Low Latency in Next-Gen Wireless Networks," International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), DOI: 10.17148/IJARCCE.2022.111258

[4] Silva, A., & Yamaguchi, T. (2024). A Web-Centric Cloud Framework for Real-Time Monitoring and Risk Prediction in Clinical Trials Using Machine Learning. Computers in Biology and Medicine, 171, 107982. https://doi.org/10.1016/j.compbiomed.2024.107982.

[5] Meda, R. (2022). Integrating Edge AI in Smart Factories: A Case Study from the Paint Manufacturing Industry. International Journal of Science and Research (IJSR), 1473–1489. https://doi.org/10.21275/ms2212142906

[6] U.S. Food and Drug Administration. (2013). *Oversight of clinical investigations—A risk-based approach to monitoring: Guidance for industry*. https://www.fda.gov/media/116754/download

[7] End-to-End Traceability and Defect Prediction in Automotive Production Using Blockchain and Machine Learning. (2022). International Journal of Engineering and Computer Science, 11(12), 25711-25732. https://doi.org/10.18535/ijecs.v11i12.4746

[8] AI-Based Financial Advisory Systems: Revolutionizing Personalized Investment Strategies. (2021). International Journal of Engineering and Computer Science, 10(12). https://doi.org/10.18535/ijecs.v10i12.4655

[9] TransCelerate BioPharma Inc. (2013). *Risk-Based Monitoring Methodology—Position Paper*. https://www.transceleratebiopharmainc.com/wp-content/uploads/2013/10/TransCelerate-RBM-Position-Paper-FINAL-30MAY2013.pdf

[10] Krishna Reddy Koppolu, H., Recharla, M., & Chakilam, C. (2022). Revolutionizing Patient Care with AI and Cloud Computing: A Framework for Scalable and Predictive Healthcare Solutions. International Journal of Science and Research (IJSR), 1457–1472. https://doi.org/10.21275/ms2212142204

[11] Barnes, S., Baird, G., Cardone, R., & Puthumana, J. (2014). Technology considerations to enable the risk-based monitoring methodology. *Therapeutic Innovation & Regulatory Science, 48*(3), 276–285. [https://doi.org/10.1177/2168479014546336]

[12] Hurley, C., Sinnott, C., Clarke, M., Kearney, P. M., Racine, E., & Eustace, J. (2016). Risk-based monitoring (RBM) tools for clinical trials. *Contemporary Clinical Trials, 51*, 15–27. https://doi.org/10.1016/j.cct.2016.10.005

[13] Goutham Kumar Sheelam, Botlagunta Preethish Nandan. (2022). Integrating AI And Data Engineering For Intelligent Semiconductor Chip Design And Optimization. Migration Letters, 19(S8), 2178–2207. Retrieved from https://migrationletters.com/index.php/ml/article/view/11913

[14] Meda, R. Enabling Sustainable Manufacturing Through AI-Optimized Supply Chains.

[15] Inala, R. Advancing Group Insurance Solutions Through Ai-Enhanced Technology Architectures And Big Data Insights.

[16] Sheelam, G. K. (2022). Reconfigurable Semiconductor Architectures For AI-Enhanced Wireless Communication Networks. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3867

[17] Buyse, M., George, S. L., Evans, S., Geller, N. L., Ranstam, J., Scherrer, B., et al. (2020). Central statistical monitoring of investigator-led clinical trials. *International Journal of Clinical Oncology, 25*(7), 1291–1300. https://doi.org/10.1007/s10147-020-01726-6

[18] Izmailova, E. S., Ellis, R., & Benko, C. (2020). Remote monitoring in clinical trials during the COVID-19 pandemic. *Clinical and Translational Science, 13*(5), 838–841. https://doi.org/10.1111/cts.12835

[19] Dwaraka Nath Kummari,. (2022). Machine Learning Approaches to Real-Time Quality Control in Automotive Assembly Lines. Mathematical Statistician and Engineering Applications, 71(4), 16801–16820. Retrieved from https://philstat.org/index.php/MSEA/article/view/2972

[20] Marra, C., Chen, J. L., Coravos, A., & Stern, A. D. (2021). Use of connected digital products in clinical research following the COVID-19 pandemic: A cross-sectional analysis. *BMJ Open, 11*(6), e047341. https://doi.org/10.1136/bmjopen-2020-047341