International Journal of Information Technology Project Management
Volume 14 • Issue 1

Mining Project Failure Indicators From Big Data Using Machine Learning Mixed Methods

Kenneth David Strang, RMIT, Australia & W3 Research, USA
https://orcid.org/0000-0002-4333-4399

Narasimha Rao Vajjhala, American University of Nigeria, Nigeria*
https://orcid.org/0000-0002-8260-2392

ABSTRACT

The literature revealed approximately 50% of IT-related projects around the world fail, which must frustrate a sponsor or decision maker since their ability to forecast success is statistically about the same as guessing with a random coin toss. Nonetheless, some project success/failure factors have been identified, but often the effect sizes were statistically negligible. A pragmatic mixed methods recursive approach was applied, using structured programming, machine learning (ML), and statistical software to mine a large data source for probable project success/failure indicators. Seven feature indicators were detected from ML, producing an accuracy of 79.9%, a recall rate of 81%, an F1 score of 0.798, and a ROCa of 0.849. A post-hoc regression model confirmed three indicators were significant with a 27% effect size. The contributions made to the body of knowledge included: A conceptual model comparing ML methods by artificial intelligence capability and research decision making goal, a mixed methods recursive pragmatic research design, application of the random forest ML technique with post hoc statistical methods, and a preliminary list of IT project failure indicators analyzed from big data.

KEYWORDS

Big Data, Information Technology, Machine Learning, Model, Prediction, Project Failure, Project Management, Random Forest

INTRODUCTION

Approximately half of Information Technology (IT)-related projects around the world have failed (Kurek, Johnson, & Mulder, 2017; Masticola, 2007; Strang, 2021). In 2009 the U.S.-based Standish Group (2009) found only 32% of projects in the American government were successful, the remaining 68% were challenged or an outright failure. In European Union countries, a 50% procurement project failure rate was discovered from the large rigorous seminal study by Ghossein, Islam, and Saliola (2018). The nearly 50% project failure rate was corroborated in two large rigorous empirical U.S.

DOI: 10.4018/IJITPM.317221 *Corresponding Author

government-based studies, with no statistical evidence found to account for the problems (Borbath, Blessner, & Olson, 2019; Eckerd & Snider, 2017). Pace (2019) argued that U.S. IT-related project failure rates have remained steady over at least 20 years despite significant advances in software and methodologies. A high project failure rate even up to 90% may be expected in industries such as R&D or space exploration, but not in IT. Israel (2012, p. 76), a former project leader at the U.S. Federal Bureau of Investigations, reviewed decades of IT-related public projects from an insider perspective, and he wrote this surreal synthesis "the federal government has wasted billions of taxpayer dollars on failed projects." This 50/50 gamble of project success vs. failure must leave stakeholders feeling perplexed about why analytical approaches have improved other fields like medical drug prediction (e.g., cancer, COVID-19, etc.), yet an IT project sponsor's ability to forecast success is statistically about the same as guessing with a random coin toss.

From a researcher perspective, it seems unusual that only a few of the project management-related journals have published empirical studies to explore the high project failure rate in an effort to improve the body of knowledge. The references illustrate which journals are championing the scientific search for this elusive answer. The authors felt it was frustrating that some journals predominately published single case studies of so-called megaprojects (i.e., large projects). Most often the goal of a single case study was to discuss a project success, not a failure, at one site. The problem with those studies was that the results were speculative and difficult to generalize, such as studying risk management at a large global oil platform in an oligopolistic market. It was not clear if the findings were statistically significant and more so any results would generalize only to equivalent populations, namely other oil rigs in the ocean. Other journals have favored surveys or interviews to collect perceptions of failure. Three problematic issues with those survey data collection approaches were poor designs, common method bias (no triangulation of evidence) and asking opinions of project performance instead of collecting actual metrics.

On the positive side, some empirical studies have revealed what is causing projects to fail. Attributes such as ISO quality approval, years of experience, prior project duration, communication skills, leadership, project manager (PM) certification, gender, corruption and incompetency — ineffective project management — have been found to impact project outcomes (Anthopoulos, Reddick, Giannakidou, & Mavridis, 2016; Jennings, Lodge, & Ryan, 2018; Laurie, Rana, & Simintiras, 2017; Martinez-Perales, Ortiz-Marcos, Ruiz, & Lazaro, 2018; Ngonda & Jowah, 2020; Pace, 2019; Saadé, Dong, & Wan, 2015; Strang, 2021). The problem with those empirical studies was the small effect sizes which means when a causal factor was identified the practical impact was negligible, leaving 88-98% variation unaccounted for. For decision makers, this means the significant models of project failure have a small economic utility as compared to the unknown factors. For other stakeholders including higher education professors, project management practitioners, and IT management associations, those small effect sizes were not enough to justify amendments to the body of knowledge.

The authors believed machine learning (ML) algorithms could advance the state-of-the-art in identifying the critical failure factors of IT-related projects. ML is a nonlinear evidence-based technique which can be used on very large datasets with many non-normal variables of mixed data types as well as missing values. Quite often the government of democratic nations will capture relevant project details and make these data available to researchers. If enough project details were available this could result in a very large sample size and possibly by applying ML, the authors could shed some light on what is causing projects to fail. Consequently, the goal of the current study was to explore new mixed methods including ML for identifying the unknown project failure indicators by using distinctly different very large retrospective project big data. Subsequently, the primary research question (RQ1) became: Can ML explain why thousands of IT projects failed by mining hundreds of big data attributes? This led to the second research question (RQ2): What were the most likely indicators associated with IT project failure? While it is acknowledged the current study may not surpass existing causal effect sizes, it is hoped that by introducing new mixed method analytical approaches with big data, this will stimulate other researchers and stakeholders to collaborate on a

mutually beneficial goal of improving the project management practice. The rationale underlying the current study was that by demonstrating an empirical method and identifying even a few IT-related project failure predictors, this would motivate practitioners and decision-makers to establish a research agenda towards improving the body of knowledge and practice.

LITERATURE REVIEW

The empirical literature was reviewed using a multi-database tool, which accessed several good-quality full-text indexes, including ProQuest Central and EBSCO. The keywords Project + machine learning, project + success + machine learning, project + failure + machine learning, project + critical success factors, and project + failure was applied in searches. The filter conditions were English, peer-reviewed journals, peer-reviewed conferences between 2007 and 2021. Originally a five year look back was used but since only a few relevant empirical results were returned (i.e., theoretical literature reviews were not desired), this period was lengthened based on trial and error using the search engines.

What We Know About IT Project Failure

Overall, IT project success/failure was a popular topic in international literature despite no conclusive answers. There were numerous studies published on the keywords project success or project failure - a quick scan of the literature in Google Scholar returned over a million papers across many disciplines and industries. Scholars have creatively tried many approaches to find the project failure causes, including collecting data from many levels (project managers, teams, organization, macro-environment) and applying standard statistical techniques (association such as correlation, regression, factor analysis, predictive linear/nonlinear modeling; as well as comparisons using t-tests, ANOVA, etc.).

Generally, the 50% project failure rate was consistent in the literature, although not every study was specifically focused on IT projects. The generally accepted criteria in the literature for project success were: Meet scope requirements, be on time, produce acceptable quality, and be at or under an agreed-upon budget (Borbath et al., 2019; Goel, 2018; Patil & Gogte, 2020). The U.S.-based Standish Group has maintained the well-known Chaos database on a longitudinal basis to investigate the causes for IT software project success and failure since 1984 (Kurek et al., 2017). The database has over 120,000 registered IT projects from over 1000 organizations (Kurek et al., 2017).

Most importantly, the Standish Group claimed that 53% of the projects were significantly over budget and behind schedule, while 18% failed outright even to the extent of causing the company to falter (Masticola, 2007). One Standish Group Chaos report stated that only 29% of the IT software projects in the U.S. were considered successful by 2004 (Masticola, 2007), and a more recent analysis found the rate rose slightly to 32% by 2009 (Standish-Group, 2009). This results in an estimated IT project failure rate of 68-71% in the U.S. Other empirical studies place the IT project failure rate between 41-50% (Anthopoulos et al., 2016; Eckerd & Snider, 2017; Ghossein et al., 2018; Strang, 2021). More specifically, the IT project success/failure factors in relevant empirical studies usually included individual and or organizational attributes. Some relevant studies used large sample sizes and robust statistical techniques, thus warranting a brief discussion. A few of these are reviewed chronologically below.

The robust empirical study by Catanio, Armstrong and Tucker (2013) was the first relevant paper in the review. They surveyed 93 project managers working in the IT industry for their self-reported perceptions of performance. Their goal was to determine how PM certification impacted the 'iron triangle' of effective project scope, time, and cost management. Although a small sample, their research design was commendable, their methods were well explained as well as executed properly (with effect size estimates reported). The unique aspect of the research design was a comparative strategy, where they used theoretical selection as coding in the survey, to separate the two groups between uncertified PMs versus certified PMs. This methodology allowed them to correctly apply comparative nonparametric and parametric statistical techniques including Chi square test of independence and

student t-tests. One slight anomaly was their data indicated the actual sample size was 87 (43 certified and 44 uncertified PMs). Despite a commendable goal and excellent research design, they did not find any significant causal predictors: "… no statistical difference between uncertified and certified project managers on their performance of project scope, time, and cost management activities" (Catanio et al., 2013, p. 158). In their discussion they acknowledged the limitations of the small sample size and self-reported perceptions of the respondents. In their discussion of the implications, they affirmed nothing was proven in that they did not find any statistically significant relationship between scope, time or cost performance and PM certification versus not being PM certified. Perhaps if 93 IT project sponsors had been surveyed and asked to provide actual project scope, time and budget metrics, along with PM certification, for a specific project, this would have yielded interesting quantitative analysis of IT project success or failure.

Saadé, Dong and Wan (2015) was another relevant and rigorous yet smaller empirical study of the project success factors, with statistically significant results. They collected self-reported data from 66 survey respondents at one United Nations agency. They were the only researchers in our review to rigorously apply exploratory factor analysis to develop a predictive model of project performance. A key rationale underlying their study was attributed to the lack of predictive project performance models in the literature. They created 19 individual-level items from a priori literature to develop the survey. Their survey questions represented skills such as communication (verbal, written) at multiple levels, ambiguity/change, escalate, attitude, cultural fit, education, leadership, prior engagement duration, past team size managed, certification, work history, technical knowledge, hands-on experience, commitment, coordination ability, situational management, and competence. They reduced the 19 to 12 items, resulting in three factors with acceptable Cronbach alpha reliabilities ranging from 0.78 to 0.95. The original three factors were engagement, education, and experience. Their discussion was confusing in several instances when they claimed they had two factors not three and using counts as well as sums instead of averages and standard deviations in the analysis, which undermined the credibility of their excellent work. However, they included the data, so we were able to calculate the estimates. The engagement factor had the highest mean which we calculated as 4.1 (SD=0.4), representing the ability to deal with ambiguity, commitment to the project, situational management skills, positive attitude, effective leadership, and effective verbal communication. Education was the next more important with a mean of 3.2 (SD=0.3), which included formal education, PM certification and writing skills. Experience had a mean of 2.7 (SD=0.08), which included length of past engagements and past team size. They dropped the work history item from the experience factor, stating it had a negligible increase on factor loading, although it can be seen that the mean was 3.3 (SD=0.8).

Saadé et al. (2015) interpreted those findings based on their experience as indicating communications and lobbying skills were the most important factors leading to project success whereas PM certification was not viewed as important in the context of the UN agency sampled. A few researchers also found PM certification did not impact project success (Nazeer & Marnewick, 2018; Pace, 2019) yet others found the opposite (Catanio et al., 2013; Crosby, 2012; Huang & Cappel, 2018). Saadé et al. (2015) did not report effect size but based on the 0.45 cumulative factor loadings of an earlier model, we could estimate the effect size would approximate 20% meaning it was a reasonably strong model. Their study had a strong research design, it was well-executed, and their limitations were honest by noting the exploratory nature and small sample size from one UN organization. They recommended their design be replicated with all factors using logistic regression.

Eckerd and Snider (2017) published a large exhaustive study using generalized least squares regression to retrospectively examine if PM attributes could predict DOD procurement program performance by collecting data from 1073 projects during 1997-2010 from the Defense Acquisition Management Information Retrieval system (DAMIR). They explained DAMIR is an executive information system operated by the Office of the Under Secretary of Defense for Acquisition. In their study, none of the contractor factors like size, experience, quality, past success, PM certification, age, or gender had significant correlation with the dependent variables of project cost variance or breach.

They noted government projects could be managed either in-house with employees or by contractor-based project managers, although they did not test those factors. They determined that aircraft, ship, and space system projects were more likely to fail in terms of a baseline breach condition, but they tended to have significantly lower cost variances. Interestingly, they determined as programs became more complex, projects tended to have less favorable cost variances although not necessarily resulting in a baseline breach condition. Thus, there was a tendency to see mixed results between project baseline breaches and cost variances, which may not be surprising given the weak beta coefficient estimates ranging between -0.78 and -0.95 they reported between the two outcomes. An interpretation of their results would be that contractor industry type was not related to project performance, although certain divisions, namely aircraft and space systems, had higher overall percentages of failures. Nevertheless, they suggested other researchers ought to test all those variables in future empirical studies.

Ghossein et al. (2018) published the largest relevant empirical study in literature. They used Pearson correlation to quantitatively examine mixed data type variables collected from 59,816 public procurement projects of European Union member countries to determine which factors were related to performance. The sample included small-to-medium-sized-enterprises (SME) and large firms. They appropriately applied a 90% level of confidence due to the exploratory nature of their study, yet they were not able to find support for most of their hypotheses. They found the age and size of firms impacted performance, but revenue was not found to be related to contractor performance. They also examined the relationship between firm size, structure, certification/training, experience (in years), exporter status, foreign ownership, access to finance, crime losses, growth rate per GDP capita, level of development, land area owned, service sector, geographic region and procurement project performance.

Ghossein et al. (2018) indicated that ISO quality status had a positive influence on product innovation, process innovation and R&D spending for manufacturing firms, but no statistically significant effect for service firms. Their data seemed to indicate that in terms of innovation, ISO quality had a far greater effect on manufacturing firms than services firms. They were one of the few researchers to address corruption as a project impact factor, claiming that effective procurement project management systems were negatively correlated with corruption faced by the business sector. However, it was unclear exactly how they measured corruption. Despite the monolithic sample size and several significant coefficients, the Ghossein et al. (2018) study had a few limitations including missing hypotheses, small effect sizes, logarithmic transformations which could obscure factor interaction, and lack of proven causality due to the correlational design.

Borbath et al. (2019) was the second largest relevant empirical study. They applied Spearman correlation to examine the performance of 14,836 contractor projects across the three army, navy and air force divisions, based on data provided by DOD Contractor Performance Assessment Reporting System. Their goal was to determine which factors were related to contractor project performance across three ordinal dependent variables: Cost, schedule, and technical quality score. They tested numerous individual and organizational level factors including age, gender, PM certification, education, experience, quality, tenure, revenue, industry type, management or business relations, firm size, and subcontracting status (in-house versus out-sourced). They argued that the longer the government program manager (PM) was in the position (e.g., tenure, experience and age), the better the decisions and program outcomes ought to be. They also argued it could make a difference if the leader was selected in-house (where better training maybe available) would be different as compared to out-sourced with the contractor. Unfortunately, none of their hypotheses were supported, but their selection of project performance predictors, honesty and sample size made this another benchmark study.

Strang and Perez (2020) examined 927 U.S. air force, navy and army supply chain projects in a rigorous quantitative study using logistic regression to identify the significant predictors of external contract outcomes. They followed the recommendations and factors identified by Ghossein et al. (2018) but added more items and they applied logistic regression instead of correlation. It appears the projects they sampled were not necessarily IT-focused (it was supply chain logistics) but the

results would likely be relevant to generalize in terms of the success or failure indicators since project management was a key factor of interest. They tested age, gender, culture, education level, experience in years, contractor ISO quality certification, and project manager professional certification in terms of whether those variables impacted procurement contract results. They found contactor ISO quality certification and PM certification were statistically significant. Those two factors could correctly classify 67% of the contract outcomes but their model explained only 11% of the variance, thus leaving 89% caused by unknown factors, and 33% of the other contracts incorrectly categorized. Although their sample size was much smaller than similar studies by Borbath et al. (2019) as well as by Eckerd and Snider (2017), it was encouraging to observe they were able to extend earlier efforts by finding several statistically significant project performance predictors.

Ngonda and Jowah (2020) published a quantitative study using correlation to assess the relationship between 306 PM's and their organization's project management maturity in South Africa. They used a survey to collect data concerning PM soft skills such as leadership, competence, power, influence, and certification along with demographics and organization project maturity level. They argued PM competence was a key factor that ought to impact an organizations' capability to successfully execute projects, which could be considered project management maturity. They did produce significant results, with positive correlations between PM power, PM technical expertise and their organizations' project management maturity level. They interpreted the findings as when a PM's power and technical expertise increased it was likely that their organization's project management maturity also increased. However, they noted the effect sizes were extremely small, and causation could not be established as it was not possible to establish the temporal order amongst the factors and dependent variable. They had a well written study, and they were honest in their limitations being the country-based population sample of self-reported perceptions and that the correlations were too weak to be used for generalization.

Strang (2021) used logistic regression in a rigorous empirical study to examine 2692 IT-related U.S.-based defense sector projects valued at or over $1M USD, which had completed in 2019. That study was similar to Borbath et al. (2019) as well as by Eckerd and Snider (2017), but higher priced projects were selected, the sample size was more than double that of Eckerd et al., and causal logistic regression was applied instead of correlation or GLS. Strang followed the recommendations of Saadé et al. (2015), by testing all 19 items they identified, along with others, but by applying logistic regression instead of exploratory factor analysis. The factors he examined after reviewing the a priori literature, included: Industry type, revenue, organization size, ISO or OSHA quality approval, gender, age, education level, years of experience, project manager certification, and he proposed maturity level, employee turnover, as well as prior success as predictors. He reported 41% of those projects were considered failures. He found PM age, PM gender, PM experience, PM certification, and organization ISO quality approval were significant in the model but accounted for only 12% of the variance in IT project outcomes, leaving 88% unfounded. He acknowledged those limitations and others, with a large conclusion subsection focused on stimulating future quantitative empirical research to uncover more of the unknown project failure factors.

Machine Learning Overview

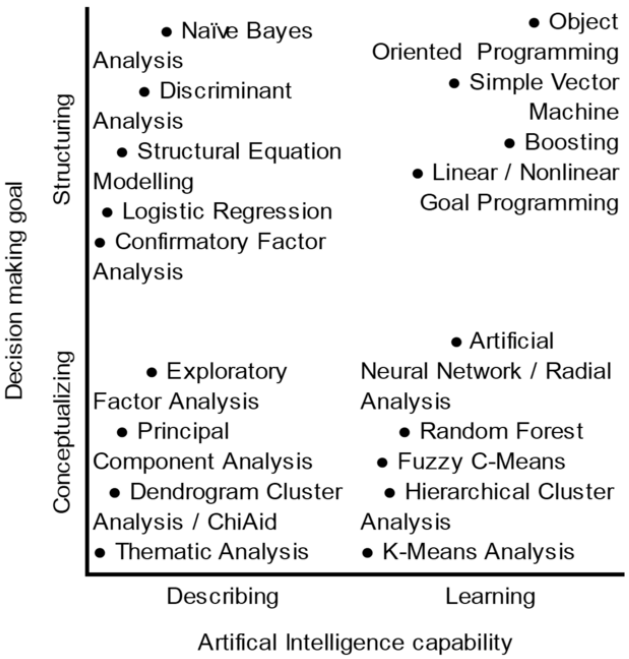
ML is a multidisciplinary and interdisciplinary field combining business, statistics, computer science, information technology, and operations research (Momoh, Rakshit, & Vajjhala, 2022). Alternatively, some researchers have defined ML as any situation where several inputs, such as raw data, are used to build a model for generating predictions and valuable insights (Hu et al., 2009). For example, programming including PHP, JavaScript, Visual Basic, C++, and object-oriented environments can be used for ML. ML is an industry 4.0 methodology that can revolutionize many fields, including project management (Toorajipour, Sohrabpour, Nazarpour, Oghazi, & Fischl, 2021). ML is considered revolutionary due to its ability to process big data and identify complex relationships to inform decision-making (Biba, Ballhysa, Vajjhala, & Mullagiri, 2010; Hu et al., 2009).

ML has already been applied in many fields and industries for pattern identification, statistical learning, human behavior prediction, data mining, speech recognition, computer vision, natural language processing, and cybersecurity intrusion detection (Wang, Fu, He, Hao, & Wu, 2020). ML is considered a relevant technique for real-time artificial intelligence (AI) analytics (Biba et al., 2010). AI is defined as the capability of machines to communicate with and imitate the capabilities of humans (Memeti, Pllana, Binotto, Kołodziej, & Brandic, 2018). The advantage of using ML for AI is that the techniques can learn relationships from data without defining all assumptions about the underlying mechanisms (Biba et al., 2010). For example, ML algorithms can model complex human problems, including face recognition, email spam filtering, speech recognition, weather forecasting, anomaly detection, churn prediction, failure detection, and document classification (Memeti et al., 2018).

The authors define ML as advanced data analytics method which may embed parametric or distribution free nonparametric statistical techniques to identify the significant factors and patterns from very large mixed type big data potentially having missing values. Big data refers to large volumes of information often scattered across many sources, beyond what can be viewed on a screen such as Excel or SPSS – often advanced SQL or Hadoop database languages are needed to query big data sources to answer fuzzy research questions or to test exploratory statistical hypotheses (Strang, 2015). The multifaceted goal of ML is to process large complex mixed data types (where no customary statistical program could function), in order to identify a subset of likely factors to simplify complex big data or to describe the patterns within the big data which are of interest to decision-makers (Oliveira, 2019). Identifying repeating patterns in big data is what gave rise to the term machine learning.

Figure 1 is a conceptual model developed by the authors which will be explained below. In this model, ML is partitioned into two broad categories based on its purpose and capability – conceptual factor description and structured predictor learning. Some techniques used in mixed-method designs, such as regression, will span more than one category (Strang & Sun, 2022). AI is beyond the scope of

Figure 1. Common ML technique typology (Adapted from: Strang & Sun, 2022)



ML because it requires additional programming to simulate behavior. Descriptive ML techniques can begin with a blank slate, allowing the ML to identify factors based on patterns and or mathematical formulas. These types of ML are often called basic analytics or clustering techniques. The underlying mathematical algorithms can include statistical mode, median or mean calculations, coefficient of variation, Euclidian distance, or similar nonparametric formula (Strang & Sun, 2022). A few ML classification techniques such as cluster analysis, exploratory factor analysis, contingency analysis, thematic node analysis, and ChiAid allow researchers to suggest labels representing conceptual groups identified from the data (Strang & Sun, 2022). This classification approach is ideal for researchers wishing to make sense of big data by superimposing business or academic terms on the proposed clusters or nodes.

Qualitative classification ML techniques may involve more researcher control, such as thematic node analysis using NVIVO or Atlas software (Strang & Sun, 2022). The classification ML approaches commonly follow either a divide up or build up principle. A divide-up approach starts by separating dissimilar data until what is left is less dissimilar to items remaining in the group than items in other groups. The buildup tactic starts with a blank slate, adding factors with some similarity into nodes such that items in each group are more similar than items in other groupings. Numerical content codes could judge the similarity, and patterns within alphanumeric fields, patterns between fields across records, abstractness, or other fuzzy segmentation concepts. Descriptive ML techniques such as principal component analysis, exploratory factor analysis, or dendrogram cluster analysis do not identify a dependent variable (Strang & Sun, 2022). Thus, prediction is not possible, although behavior patterns may be identified. These ML techniques commonly have minimal assumptions and accept any data type except graphical fields. Thus, text phrases captured from online blogs or cell phone taps can be analyzed and classified into behavior patterns using ML. The more advanced structured ML techniques require conditional programming to re-process big data and make recommendations according to generated estimates and retraining, which is the precursor to AI learning.

The ML field is very complicated, so the authors developed the conceptual model to summarize the generally accepted approaches from a decision-making perspective. Figure 1 illustrates our conceptual taxonomy of the family of ML techniques decision-makers use. We include only commonly used ML techniques (it is not an exhaustive list). In a few cases, for example, logistic regression, there are numerous variants, including ordinal, probit, and multiple regression techniques also used as part of ML applications - our diagram serves to depict where we feel each ML technique falls according to decision-making goal and AI learning capability. The bottom x-axis in Figure 1 represents an increasing ability of ML to operationalize AI learning. Starting first on the left, we have ML techniques known for describing, classifying, or structuring complex big data. On the right, we show ML techniques that have dependent variables to predict or learn behavior. Learning for AI is possible only with conditional programming where go-to and other recursive search procedures are possible.

The y-axis of figure 1 represents the decision-making goal when an ML technique is deployed. The lower quadrants on the y-axis are meant to illustrate conceptual ML techniques that decision-makers use early in the process when there are many unknowns. Big data is unstructured, and the goal is to begin to make sense of it by identifying similar independent factors or market segmentation groups. In the conceptual phase, ML techniques could have dependent variables, but the confidence or reliability of the result is not known or not easily generalizable to future problems. By contrast, the structuring quadrants of the y-axis in Figure 1 contain what we argue are advanced ML techniques with which estimate reliability, effectiveness, validity, accuracy, or significance, such as for a hypothesis test using seeking a p-value below a significance level. Lower values on the y-axis in Figure 1 refer to ML techniques that provide decision-makers with abductive reasoning. Abductive is different from inductive because the latter creates concepts without statistical confirmation estimates. That model may explain the data relationships with preliminary validation estimates but is not proven.

By contrast, higher values on the y-axis in Figure 1 depict ML techniques that are deductive. The findings contain behavior prediction effectiveness, reliability, or significance estimates to generalize

the results to similar populations. There is a parallel notion between nonparametric statistical techniques and the abduction quadrants in Figure 1, particularly the descriptive-conceptual ML quadrant. Since distribution-free procedures are less empirical, they often lack significant estimates. Similarly, there is an inherent similarity between parametric statistical techniques and the deduction quadrants of figure 1. Logistic regression is a common parametric statistical technique for testing quantitative scientific hypotheses. A key attribute of ML techniques in the top right deductive-predictive quadrant is conditional programming, the ability to recourse the data (parse it more than once), make nonlinear go-to decisions based on information assembled in arrays from previous record analysis. This latter feature provides better AI learning capability beyond merely predictive decision trees or factor reliabilities.

To further explain Figure 1, we propose a logical grouping of the ML techniques into the four quadrants. Starting at the lower left, we have conceptual-descriptive ML techniques, such as cluster dendrograms, principal component analysis, and exploratory factor analysis, classifying big data for subsequent analysis. Some decision-making may be possible at this phase, particularly in marketing, where consumer buying behavior is analyzed to create product design and advertising segmentation. The structured-descriptive ML quadrant in the upper left represents advanced techniques, including structural equation modeling (SEM), logistic regression, discriminate analysis, naïve Bayes analysis, confirmatory factor analysis, and other methods which require numeric data types and generate significance estimates along with effect sizes. The bottom right quadrant represents predictive ML techniques with recursive routines to test or train a decision-making process necessary to support AI in a larger programming context.

Common conceptual-learning ML techniques positioned in the bottom right quadrant of figure 1 include random forest, artificial neural network/radial analysis, fuzzy c-means, hierarchical cluster analysis, and K-means analysis. The top right quadrant contains AI-capable ML techniques, which can predict behavior by making conditional branches in logic based on recursive big data analysis. These are structured-learning ML techniques. Video games commonly employ AI through object-oriented programming (OOP) languages, including C++. Support vector machines (SVM) or boosting are typically applied for big data analysis in the structured decision-making context. Still, linear or nonlinear goal programming can also be operationalized in simple Excel spreadsheets if condensed big data is available.

Applying Machine Learning to IT Project Failure

Although there were few studies where researchers applied ML to identify project failure factors, several applied ML techniques to analyze IT project scheduling decisions to avoid failure. There are two dominant ML approaches in the literature for planning IT software projects' time or schedule aspect: Static and dynamic models (Fatima et al., 2020; Memeti et al., 2018). Static models employ a retrospective ideology, data is used to develop or explain the duration, but the project schedule (or activity path) does not change (Pospieszny et al., 2018). In the dynamic model, data are used to quantify unknown risks and predict progress, while task estimates, duration and scheduling may be improved (Fatima et al., 2020). In some empirical studies, ML has been successfully applied in the dynamic model where there were scheduling changes because of uncertainty (Fatima et al., 2020). Alternatively, ML data mining techniques have been found to be highly significant for quantifying uncertainty and predicting effort/duration estimation in the initial stages of the IT software project lifecycle (Patel, Modi, & Sarvakar, 2014; Pospieszny et al., 2018).

A few researchers argued that ML techniques are better than traditional decision-making approaches for project risk management because they reduce bias. Applying ML algorithms using historical big data as input reduces the impact of human bias, which we argue may account for the inability of current researchers to conclusively identify the IT project failure indicators. There are three categories of biases that influence traditional decision-making techniques: Technical, psychological, and political (Pospieszny et al., 2018). The technical bias results from estimates based on imperfect

information, while the psychological and political biases are based on the human factor and the inability to plan appropriately (Pospieszny et al., 2018). The advantage of using ML algorithms over statistical and mathematical algorithms is that the former can learn and improve by comparing the predicted model to observed data (Pospieszny et al., 2018). Advanced ML techniques such as Simple Vector Machines (SVM), Artificial Neural Networks (ANN), and random forest leverage supervised learning, a series of labeled samples set aside as training data and then compared to the predictive model developed using the remaining records. In the supervisory learning ML algorithm, parametric statistical techniques are applied to generate models because the goal is to make inferences from the training sample to predict unobserved future behavior. Hence, the efficiency of the training algorithm is as important as the accuracy of its factor classification (Hu et al., 2009).

ML was also applied to predict IT software defects, which could cause a project to fail. Han, Lung and Ajila (2016) applied SVM, ANN, and random forest ML techniques to identify the important code and process metrics from 102,675 IT software projects to predict defects. The most interesting aspect of their study was how they sourced big data which was six files of metrics from the open-source Eclipse project (www.eclipse.org). In this crowdsourcing initiative, programmers worldwide collaborated on three releases of an IT software program. The big data included at least 61 mixed data type variables of interest. Their research design was similar to ours in that they identified a binary dependent variable indicating whether an IT software program component was defective (1) or not (0). In the current study, our RQ focuses on whether an IT project is a failure (1) or not (0).

As with many other ML researchers, the approach applied by Han et al. (2016) was to use several techniques in advance of ML, to identify or confirm the critical indicators and dependent variable(s). They used the three ML techniques to compare them to identify the best fit for IT software project defect-failure analysis. This is one major difference between ML and non-ML studies. The big data factor relationships are typically ambiguous without any *a priori* theoretical models to hypothesize from – thus, when starting with a blank slate, it helps to know the more likely independent indicators and dependent variables, which can be validated in ML. They applied principal component analysis (nonparametric) and generalized linear model (parametric) statistical techniques on the key indicators with the proposed continuous dependent variable (defect) in a reduced sample to confirm there was some correlation and predictive ability in the big data.

As they stated, these preliminary steps were not required in ML. Instead, they explained it was done to save computational time since they were confident, they could specify the key indicators and one dependent variable, thus narrowing the choice down from 61 to 27 predictive features. They applied three ML techniques, CVM, ANN, and random forest. After applying ML, they interpreted the confusion matrix and Receiver Operating Characteristic (ROC) curve estimates to identify a significant result. They found ANN and random forest ML were preferable to SVM for experimental samples. The ROC curve for ANN was 0.878, 0.844 for random forest, and 0.782 for SVM. On the other hand, they reported the prediction accuracy was slightly higher for SVM with a sensitivity estimate of 0.7988 compared to 0.7778 for ANN and 0.7717 for a random forest. Thus, the selection of ML technique will depend not only on the RQ and hypothesis, but also on the data, particularly if a dependent variable is available and what type.

METHODS

In the current study, the authors adopted a pragmatic research design ideology. A pragmatic ideology refers to researcher's intention to focus on factual evidence was sought to prove deductive theories, often quantitative data types are preferred to facilitate analysis, much like a post-positivist philosophy, but practical approaches are permitted to achieve the goals. In a pragmatic ideology, formal methods are often customized and mixed with atypical procedures, to overcome a significant constraint (Strang, 2015). A recursive pragmatic research design was applied, including several ML techniques, a structured object-oriented programming language, and a spreadsheet was sometimes used specially

to refine the analytic charts, and a commercial statistics program (SPSS v. 25) was used for parametric analysis. The recursive aspect of the design meant iterations of structured programming took place forcing a replication of the ML and statistical techniques. A pragmatic ideology was also applied to the literature review, using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) technique. The authors applied PRISMA by continually eliminating articles in the search results if they were not empirical and related to project outcomes/performance. Excel was used to weigh each article and sort the higher priority papers, to provide the final short list for discussion in the current paper.

Ethics and Sample Data

The authors obtained approval from their employers to conduct the study. The first author (Strang) was the principal investigator (PI). The PI designed the study, wrote the initial paper, conducted the analysis including programming and interpreted the results. The PI obtained ethical clearance to conduct the study from the internal research review board. The PI obtained the large big data source link from the U.S. Government Accountability Office (USGAO, 2020). The second author (Vajjhala) served as the corresponding author, and he presented the conceptual research design to peers at an international information systems big data conference in Europe to receive constructive feedback to refine the study. Both authors collaborated equally in the literature review and paper revisions.

The big data included close to a million records with hundreds of metrics from projects, in nominal, interval, and continuous data type formats. The data was theoretically filtered for IT-related projects in order to make the information compatible with key a priori studies. The cutoff date of 12/31/2019 was applied to avoid examining projects which may have been adversely impacted by the covid-19 pandemic, which could skew the indicators and results. The resulting data contained technology-related projects, that is, IT-related projects. The public project information was considered appropriate to answer the RQs since it would be difficult to obtain such relevant big data elsewhere. The U.S. is the third most populous country in the world with the highest public procurement expenditures of all OECD nations, at US$1,694bn for 2018 (Strang & Perez, 2020). This makes the selected project big data relevant for exploring the RQ.

Objected-Oriented Visual Basic (OOVB) was applied to fix numerous problems with the big data, including corrupted or undefined variable length fields. OOVB is a structured object-oriented programming language with basic statistical functions, similar to C++. Unlike ML or statistical programs, OOVB could be used on the raw big data to count records, count the number of corrupted or null values of a field for all re cords, and perform basic statistical estimates like average, standard deviation as long as the field values were not corrupted. Corrupted fields could be skipped by copying only valid records to a new working dataset in a cloud big database. OOVB is easier than R for basic processing of big data due to its error control functions and ability to continue processing records despite invalid field values or improperly delimited variables. Even R has difficult processing variable length fields if they are not in comma separated values (CSV) format. Whenever problems were encountered with ML, OOVB would be used to trouble shoot and fix the data, either by copying to another working big dataset or setting the field to a null empty value so it would not be counted or processed in statistical functions.

The data reflected military IT-related project attributes and metrics, with several indicators of success or failure. The categories of the projects in the data included for example, cyber security risk assessment and monitoring systems, inventory planning and procurement systems, vehicle maintenance and failure assessment, border crossing and tariff systems, property tax assessment and billing systems, food and drug safety systems, tax revenue collection and auditing systems, building code and safety compliance, education and accreditation evaluation systems, medicare health records and billing systems, traffic analysis and planning systems, railroad routing and analysis systems, air defense systems, office of project management assessment systems, water vessel registration and tracking systems, flight tracking and navigation systems, high altitude intrusion detection systems,

and others. It appeared that the data pertained to independent contractors not government employees. This was understandable since the majority of public projects are posted on procurement systems to obtain competitive bids, and the contractor with the best proposal would typically be awarded the project on a fixed cost, fixed time basis with allowance for approved change orders. Once the data were examined a workable dataset was constructed in a cloud repository. The sample size was estimated at 17,430 based on counting the carriage return/linefeed characters to signify a record end. There were several variables in the data representing project outcome as successful or a failure (e.g., a breach), which had previously been condensed into another nominal dichotomous indicator variable as a breach/failure = yes or a success/no breach = no.

Procedures

Based on the literature review, the authors were aware of priori factors related to project failure. However, the data did not contain all of those, and it contained many more. Nevertheless, as noted above, the data did contain a dependent variable. The RQs both suggested a predictive research design strategy. The RQ was beyond basic descriptive analysis, instead pointing towards predicting or learning. This narrowed the selection of ML techniques to the conceptual-learning quadrant in the bottom right quadrant of figure 1, referring to random forest, artificial neural network/radial analysis, fuzzy c-means, hierarchical cluster analysis, and K-means analysis. However, fuzzy c-means do not necessarily accommodate a dependent variable. Prior studies of software IT project failure had recommended SVM, ANN or random forest (RF) due to their high accuracy (generally above 80%) and high effect sizes (also above 80%). The authors also had a somewhat similar ML study of project software failures to use as a guide and for effect benchmarks namely Han et al. (2016).

In their study Han et al. (2016) recommended ANN, RF and SVM as ML techniques for exploring ambiguous big data because they were the most accurate, robust and reliable. RF was considered the most relevant ML technique for the current study since SVM and ANN were beyond what was needed, in terms of advanced AI learning. ANN is somewhat similar in nature to structural equation modeling (SEM) whereby SEM develops hidden latent constructs as indicators, indirect moderators or output variables. The hidden layer in ANN represents an additional complexity more suitable for modeling human cognitive behavior as compared to the team driven project attributes. Additionally, there was some concern SVM and ANN would not perform as well or even at all with mixed data types and missing values.

In accordance with the pragmatic ideology and ML practices, the authors proceeded with the RF analysis without specifying the hypothesis, according to the current practice in computer science. The approach would be to apply several iterations of RF by selecting all fields as indicators and designating the condensed binary outcome field as the dependent variable. Additional statistical techniques may become necessary according to what the results were, such as using regression in SPSS on a reduced extract in order to fully answer the two RQ's:

RQ1: Can ML explain why thousands of IT projects failed by mining hundreds of big data attributes?
RQ2: What were the most likely indicators associated with IT project failure?

RESULTS AND DISCUSSION

Preliminary Analysis

The sample big data were problematic because of numerous alphanumeric key words instead of numbers and missing or null values. The authors cleaned the data using a Visual Basic program to eliminate projects with more than 50% corrupted or null values using the carriage return/line feed character as the record delimiter in the big data file link. At the organizational level of analysis, the average number of employees was 32.3 (SD=22.7), the mean annual revenue before taxes was

$1,956,384.07 USD (SD=36,709.17), and all contractors in the sample database were limited liability corporations (note SD refers to standard deviation). According to government procurement project classifications, 35.2% were defense/security/space, 33.8% consisted of infrastructure/transportation/telecommunications, 13.7% focused on education, 9.2% were in healthcare, 5.9% covered the energy section, and 2.1% were in other industries. Slightly more than half of the contractors (55%) were registered as ISO or equivalent quality status (manufacturing or service).

At the individual level of analysis, the mean age was 41.9 years (SD=13.3) and in terms of demographic characteristics, 94% were male with the remaining 6% female. Obviously, the high ratio of males and corresponding low percentage of females would cause statistical skew of the model. On average PM's had 14.1 years of relevant experience (SD=5.9). Regarding education level, 95% had a college degree consisting of a vocational diploma, associate, or bachelor. A few had a master's or doctorate, while some had only grade school. 55.6% had achieved at least a bachelor's degree, 24.4% possessed a vocational or trade school diploma, 10.9% held an associate degree, but 4.4% reported only grade school. Most PMs, at 74%, reported being certified.

ML Analysis

Table 1 lists the essential factor classification estimates from the random forest ML technique. The records were divided into training (45%), validation (12%) and test samples (14%), which sum to the sample size. Overall, 94 nodes were formed with no more than 4 predictors per split. The test validation accuracy was 78.4%; the confirmation accuracy was 78.4% with an out-of-the-bag estimate of 50%. These results were comparable to a similar study by Han et al. (2016). However, they contrasted three ML techniques, SVM, ANN and RF using a different type of big data, software defect metrics with a failure indicator which is comparable to the data in the current study reflecting project attribute and metrics with a breach (failure) indicator. The out-of-the-bag (OOB) accuracy shows a projected estimate using observed data. Figure 2 illustrates a plot of the OOB estimates. The OOB plot visually

Table 1. Random forest ML classification estimates

Trees	Predictors per split	Validation Accuracy	Test Accuracy	OOB Accuracy
94	4	0.784	0.810	0.496

Note: The model is optimized with respect to the out-of-bag accuracy.

Figure 2. Random forest ML out-of-bag classification accuracy plot

confirms the training versus testing results were almost identical, with 72% accuracy reached early in the training process, below 10 trees. The goal of RF, similar to cluster analysis, would be to have fewer tree nodes with high accuracy.

Table 2 summarizes the significant confusion matrix estimates from the random forest ML model testing process. Here we can focus on the joint frequencies at the intersections of the observed versus predicted cells to calculate the marginal and conditional probabilities of recall sensitivity, specificity, accuracy, and precision. The No/No cell 68.1% is the true negative (TN) while the Yes/Yes cell 12.9% is the true positive (TP). The observed Yes and predicted No cell 13.5% is a false negative (FN) while the observed No and predicted Yes cell 5.5% is a false positive (FP). The recall sensitivity rate is calculated as TP / TP + FN. The precision is estimated by TP / TP + FP. The specificity rate is calculated as TN / TN + FP. The accuracy is derived from TP + TN / TP + TN + FP + FN.

Table 3 illustrates the quality of classification evaluation metrics calculated by the random forest ML technique – these were the overall rates as contrasted to our joint and marginal probabilities we manually calculated above. The last row of table 3, the average/total estimates, could be used as comparisons to benchmarks and other published ML studies. For comparison, RF estimates from the Han, Lung and Ajila (2016, p. 38) study were 0.7717 for sensitivity (precision), 0.8634 for specificity (recall), and 0.844 for ROCa. In the current study, the overall precision was 80% and the overall recall probability was 81%. Both of these estimates were comparable to the RF results in the Han, Lung and Ajila (2016) study of IT software project defects, with the current study having higher precision but lower recall. The F1 score in table 3 refers to the weighted average of the recall sensitivity and precision probabilities, which at 80% for the total line was quite good. The F1 estimate could be thought of as an effect size, similar to the adjusted coefficient of determination r^2 in a statistically significant multiple regression model. The F1 was not reported by Han et al. (2016).

The support column in table 3 shows the sub sample size for the marginal cells. Here we can concentrate on the Receiver Operating Characteristic area (ROCa) which estimates the quality of the ML factor classifications, in other words, how much accuracy in the dependent variable was represented by the factor classifications. To better understand ROCa, the reader could think of a normal distribution curve in terms of most of the data points being captured around the population mean area delimited by 80% control intervals. The ROCa was 85% (rounded) for all evaluation results. These results are good and they were almost identical to the random forest ML precision reported by Han et al. (2016) for IT software project defects, with the current study having a slightly higher ROCa.

Table 2. Random forest confusion matrix (in percentages)

		Predicted	
		No	**Yes**
Observed	**No**	68.1%	5.5%
	Yes	13.5%	12.9%

Table 3. Random forest quality of classification metrics

	Precision	**Recall**	**F1 Score**	**ROCa**
No	0.834	0.926	0.878	0.848
Yes	0.702	0.488	0.576	0.849
Average / Total	0.799	0.810	0.798	0.849

Figure 3 depicts a plot of the ROCa using the random forest ML evaluation data estimates. In the figure the breach condition refers to project failure – a no means the project was not breached, successful.

This illustrates the true positive and false negative estimates in a plot. It illustrates that both conditions were accurately predicted up to approximately the 80% true positive level with approximately a 20% false positive rate. You can think of this as approximating the Pareto 80/20 principle from Lean Six Sigma being that in general 80% of the success/failure outcomes can be traced to 20% of the causal indicators.

Table 4 lists the top 20 or so features - indicators - from the RF, sorted first by total increase in node purity (importance) and then by mean decrease in accuracy. The field description was also added, noting most fields were alphanumeric or nominal with only a few variables containing pure numbers. In the rightmost column of table 4, the mean decrease in accuracy is the amount of accuracy decrease when the factor is removed from the full model. Generally, the increase in node purity is considered the importance of the factor in the ML model. Node purity in ML is like change in r^2 effect size when a new factor is added to a linear regression model. Here we argue from experience and based on multiple regression effect size logic (O'Boyle, Banks, Carter, Walter, & Yuan, 2019), that a node purity increase of 0.02 is weak yet significant to discuss, 0.1 is moderate and 0.15 or more is strong.

In the current study we argue feature indicators with node purity at or above 0.02 are ideal for further post-hoc analysis if the average decrease in accuracy with the factor removed is positive, when there are numerous indicators, and the dataset is large or a big data source. In essence, we suggest the mean decrease in accuracy be considered simultaneously with increase in node purity, analogous to evaluating the number of indicators in a parametric regression model against the adjusted increase in effect size. Here we posit that a cutoff point of significant indicators can be signaled when both estimates are similar, the mean decrease in accuracy is positive and the increase in node purity drops below 0.02.

From table 4, PM Experience node purity was 0.063 making it the most influential factor in the model. Close to that was the Contract (Y value) or in-house (N value) factor with a node purity of 0.058, the second strongest factor in the model. This factor represented whether the PM was hired on a contract basis or was a full-time in-house employee. These were the two most important indicators in the ML model. Line of Business had a node purity of 0.048. This factor represented whether the IT project was focused on a core product (like a new software application) or alternatively a cross-functional project such as billing or human resources. The 'Remote allowed' variable had a node purity of 0.032 which represented whether the PM was permitted to work remotely. The next two

Figure 3. Random forest ML ROC area curves plot

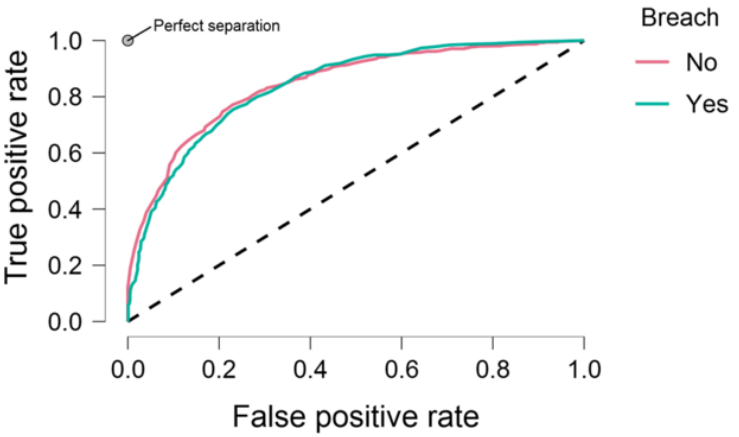


Table 4. Project failure feature (indicator) importance from random forest

Factor	Field description	Mean decrease in accuracy	Total increase in node purity
PM Experience	Years numeric	0.037	0.063
Contract or in-house	One year, two-year, N is in-house employee	0.007	0.058
Line of Business	Refers to defense industry coding	0.010	0.048
Remote allowed	Y means PM could work remote, telecommute	0.010	0.032
Contract K	$ the PM's salary or contract if outsourced	0.020	0.031
Budget K	$ baseline sponsor budget for project	0.038	0.025
PM Certified	Y if PM certified APM, Prince, PMI, Agile, etc.	0.009	0.024
**********	Authors suggested cut-off point for post-hoc analysis		
Project Quality Focus	Y ISO 9000 or OSHA registered, quality oversight	-0.001	0.015
Timesheets Captured	If earned value applied, team uses timesheets	0.0008	0.012
Cross Industry	Government or DOD partnership coding	-0.0005	0.008
Online PMO	If an online project management office used	0.008	0.006
Team Online	Y means team members could telecommute	0.002	0.006
Dependents	Code linked to other critical path projects	-1.241e-4	0.002
Prior Success	Y if same team completed 1 or more projects	6.587e-5	0.002
Training	Y if training given to team for current project	0.004	0.002
Education	Ordinal, education level of PM	0.002	0.001
Other Certs	Alpha code for other PM certification	7.757e-4	0.001
PM software used	Y if PM used commercial PM software	0.009	9.346e-4
Gender	Code for PM gender, M or F	-5.742e-5	1.725e-4
Other fields…			

fields represented the contract value and the project budget, in thousands. Contract K was the PM contract or salary amount, resulting in a node purity of 0.031, while the project Budget K had a node purity of 0.025. The PM certified field had a node purity of 0.024. Based on our established cutoff, we would stop at the above 7 fields. Nonetheless we can explain that Project Quality Focus referred to whether the organization was ISO 9000 registered, which we noted several researchers found significant in comparable studies (Ghossein et al., 2018; Strang, 2021) - it had a node purity of 0.015 putting it close to the cutoff. Finally, the Timesheets Captured field was also below the cutoff, it indicated if the organization applied earned value by capturing team timesheets, which had a node purity of 0.012 but a negligible yet positive decrease in accuracy when removed from the model. The

remaining fields contributed very little to the ML model but could possibly be of interest as controls or predictors in future studies. Based on the above interpretations, we can successfully answer RQ1, yes ML can explain why thousands of IT projects failed by mining hundreds of big data attributes, and namely we have 7 important fields with RF accuracy of 80%, a recall of 81% and an approximate F1 effect size of 80%.

In order to fully answer RQ2: What were the most likely indicators associated with IT project failure, a logistic regression was performed on an extract of just the above 7 features. First, the breach outcome indicator was converted to a numeric 1 (yes failure) or 2 (no success) because the nominal data type could not be processed. Next the continuous data types (experience, contract and budget) were positioned as covariates, with the remaining 4 as free factors, regressed on the binary breach dependent variable. A statistically significant model was produced, assuming a 95% level of confidence, with an AIC = 67.068, a BIC = 87.068, an $x^2 = 76.287$ ($p < .001$), producing an r^2 effect size of 27% (McFadden $r^2 = 0.267$).

The logistic regression confusion matrix is shown on the right side of table 5, which is directly comparable to the RF confusion matrix of table 2. To facilitate this comparison, the RF ML matrix was replicated on the left of table 5, with the logistic regression values on the right. The values are not significantly different. This can be interpreted as a good regression model with a strong effect size. It would have been very difficult to determine which indicators were likely predictors of IT project performance without first performing the ML RF on all the big data.

As a further analytic comparison of RF ML and the regression model, to illustrate the accuracy of RF, table 6 contains the confusion matrix quality of classification metrics for RF (from table 3) and those metrics from the logistic regression. Note however the RF metrics included the full sample size and all thousands of fields while the logistic regression contained only the selected 7 fields. The logistic regression program provided two additional estimates, Brier score and H-measure, which are relative indicators of factor to predictor capability ratios.

In table 6, AUC refers to area under the curve, which is a synonym for ROCa, since the statistical software uses slightly different terminology as compared to the RF ML program. The ROCa was

Table 5. Confusion matrix comparing random forest ML (all fields) with logistic regression

		Random Forest ML				Logistic Regression	
		Predicted				Predicted	
		No	Yes			0	1
Observed	No	68.1%	5.5%	Observed	0	65.3%	8.1%
	Yes	13.5%	12.9%		1	12.6%	14.0%

Table 6. Random forest (full sample) vs. logistic regression quality of classification

Measure	Random Forest	Logistic Regression
ROCa (AUC)	0.849	0.840
Sensitivity	0.810	0.526
Specificity	0.800	0.889
Precision	0.799	0.632
F1-measure	0.798	0.575
Brier score		0.138
H-measure		0.374

similar, 84% from RF versus 84% from regression. However, the classification accuracy sensitivity dropped from 81% in RF to 53% in regression, which was likely due to the processing of nominal fields by the regression algorithm (all codes were automatically converted to numbers before regressing). The specificity increased from 80% in RF to 89% in regression, likely owing to the positioning of the experience, budgetK and contractK as covariates. The precision accuracy also dropped from 80% in RF to 63% in regression. Again, this was likely due to the data types and lack of equivalent processing capability in regression for missing values or nominal codes. Finally, the F1 quality effect size of 80% in RF dropped to 58% in regression. Unfortunately, we are left without comparative importance scores in the logistic regression model due to the irregularity of the big data. That was where the RF ML technique excelled, being able to process non-normal data quite accurately.

Next, descriptive coefficients from the logistic regression were calculated for the 7 features as shown in table 7 to fully answer RQ2. Forward regression was applied, and the final best model was summarized for table 7. The standardized coefficients (beta) can help explain how the 7 features relate to the project outcome of success or failure (breach). The quantitative fields of experience, contractK and budgetK will be most meaningful since the dependent variable breach is coded as 1=no breach/success, 2=breach/failure. In other words, for this type of dependent variable coding, a negative beta is desirable to indicate if a quantitative factor has a high value, then it leads to a successful project outcome. For the quantitative covariates in table 6, the logistic regression coefficients can be interpreted as higher values for the breach code of 2 are better, and vice-versa. Interpretation of the factors is as best speculative since they are nominal. The feature remote allowed was not accepted in the regression model as it contained too many missing or null values. In table 7, the (1) and (2) in parenthesis refer to the breach conditions, SE refers to standard error. The z is a z-score with an associated Wald association estimate and the last column is the p value. Any indicator with a p value above .05 would be considered a non-significant factor so it should not be further analyzed. The standardized betas are better for factor-to-factor comparisons as they are normalized based on the distribution of values to be units comparable.

Clearly the ContractK (PM salary) and PM Certification indicators in table 7 were not significant, so they did not relate to the project failure/success outcome in the sample. PM Experience was the most important indicator, with a coefficient of -0.059, (SE = 0.006), beta = -1.441, z = -9.434, Wald = 88.996, $p < .001$ (significant). The large standardized negative beta of -1.4 could be interpreted as less experience results in a higher outcome condition of 2 (failure), and vice-versa, higher years of experience tends to end with a successful project condition (1). Project BudgetK contained missing values so not all estimates were available, yet the beta was 0.744, with a z = 4.653, Wald = 21.651

Table 7. Logistic regression predictive indicator estimates of 7 ML features

Construct	Coefficient	SE	Beta	Z	Wald	P
(Intercept)	-0.841	0.209	-2.238	-4.023	16.185	<.001
PM Experience	-0.059	0.006	-1.441	-9.434	88.996	<.001
PM ContractK	-0.004	0.003	-0.113	-1.086	1.179	0.278
Project BudgetK	0	0	0.744	4.653	21.651	<.001
Outsourced/in-house (1)	0.869	0.075	0.869	11.629	135.224	<.001
Outsourced/in-house (2)	0.008	0.077	0.008	0.099	0.01	0.921
LineOfBusiness (1)	-0.003	0.054	-0.003	-0.058	0.003	0.954
LineOfBusiness (2)	1.101	0.114	1.101	9.618	92.513	<.001
Certification (1)	0.018	0.034	0.018	0.537	0.288	0.591
Remote allowed	na	na	na	na	na	na

and $p < .001$ (significant). This can be understood as projects with higher budgets were more likely to result in a breach failure condition. Additionally, it was half as important as experience by looking at the standardized beta. Outsourced/In-house (PM was contracted or an employee) was significant only with a successful outcome ($p < .001$) with a coefficient of 0.869 (SE = 0.075), a beta = 0.869, z = 11.629, Wald = 135. Outsourced, for successful projects, had about the same importance weight as budget. This can be thought of as for projects we know failed, the PM was likely in-house (not outsourced). The LineOfBusiness was an internal defense industry coding which cannot be further interpreted, although we can deduce the 1.101 beta for non-breached projects ($p < .001$) suggests that some categories of work were more successful than others. Based on these results we can now fully answer RQ2 that we know which indicators.

CONCLUSION AND RECOMMENDATIONS

Turning back to our rationale for initiating this study, we can now answer our RQ: Can the random forest ML technique identify the most important indicators associated with IT project failure based on declassified big data? We concluded yes; the ML random forest technique was effective. Overall, the random forest ML technique was accurate and practical for analyzing big data from over 17,000 projects to identify the critical failure indicators. The ML model training accuracy was 78.4% and the validation accuracy based on the confusion matrix was 81%. Another key ML model statistic was the average area under the ROCa which was 85%. These estimates were comparable to similar studies including the IT software project defect analysis published by Han et al. (2016).

A post-hoc logistic regression model produced a 27% effect size and revealed the most important indicator was PM experience, more experience suggested avoiding a project failure. We also know that projects with higher budgets were more likely to result in a breach (failure) condition, but this was half the importance weight as experience. Outsourced/in-house was similar to budget in importance and it could be interpreted as in-house PMs (employees not outsourced) were more likely to have breached projects. LineOfBusiness was an internal defense industry classified code, and we know some types of projects were more likely to be successful than others, but we do not know if it impacted failed projects (that condition was insignificant). We also know that PM salary and PM certification had no impact on the project breach/success outcome, but we were unable to test remote work allowed.

Nevertheless, although we concluded the random forest ML was useful for initially identifying important features from big data, and we calculated their importance along with the model effect size, the authors would recommend additional statistical techniques be applied as a next step to extend this research. The authors also had to use OOVB several times to fix problems with the big data. Furthermore, in contrast to the recommendations of Han, Lung and Ajila (2016), we did not conduct preliminary analysis using a general linear model (GLM) to identify the most likely indicators because we know GLM would not accept the partially corrupt big data whereas ML would accomplish that step more effectively. The authors demonstrated a difference mixed methods sequence, with ML first followed by predictive logistic regression and post-hoc analysis, to identify the most important indicators related to IT project failure. The logistic regression corroborated that 3 of the top 7 fields were relatively good predictive indicators of the dependent variable project failure.

While the random forest ML technique performed well for this big data sample with 81% accuracy, we would encourage researchers to try other ML techniques in the structured-learning category. Other structured AI learning ML techniques would be capable of predicting behavior and generating reliability statistical estimates for the model. However, the source big data would likely have to be cleaned to ensure the fields were not corrupted and contained relevant codes or numbers (preferably). In the current study, a significant amount of effort went into cleaning and reprogramming the big data with OOVB prior to RF.

IMPLICATIONS AND FUTURE RESEARCH RECOMMENDATIONS

The authors acknowledge that a high project failure rate is not uncommon in certain industries including R&D, space exploration and others. However, as Israel (2012) pointed out, any IT project failure can be costly to direct and indirect stakeholders. Despite the current study being able to successfully answer both RQs, we still do not know enough about what is causing caused half of all projects around the world to fail over the last 5 decades. More research is needed to accomplish that immense mandate. The authors reflected on this mandate and developed several suggested implications.

When looking at the future, two recommendations arise from the current study and author reflections. Both authors are licensed project managers with a combined 50 years of experience, and they have published more than 300 empirical studies of teaching as well as project management. The first recommendation would be to replicate as well as extend the current study, with additional empirical project outcome evidence data (not surveys). We also recommend using alternative machine learning methods along with traditional parametric statistical techniques including multiple regression, for the purpose of identifying and validating the critical project failure indicators.

We do not recommend using surveys to collect data because we assert that asking for opinions or perceptions of project failure indicators will not be accurate since they are too far after the actual failure events, and opinions are never as good as actual empirical indicators, such as the project metrics. Additionally, the authors assert that a PM, a team member, even a sponsor will be impacted by conscious or unconscious bias when asked to grade a project they were a member of. Bias could impact the respondent in two ways, first causing them to overestimate and inflate their memory of success, or secondly, answer with negative sentiments due to having a bad experience connected to the project. Bias is difficult to measure and filter out, so it is better to not ask for it. We are not saying that surveys are terrible data collection techniques or that project member opinions should never be sought. No, we acknowledge surveys are ideal especially for collecting qualitative opinions of complex project impact factors. Instead, we specifically point out that IT project performance data would likely already exist in organizations that practice project management, and actual data are a better source of evidence, in that circumstance, as compared to using surveys to collect opinions of metrics long after the project finished. The challenge we found though is accessing the data and then transforming or correcting messy unstructured big data to accommodate statistical analysis. This is where the authors have proven that a pragmatic ideology with mixed methods including ML and OOVB programming with statistical modules can be valuable. A pragmatic ideology and mixed methods with ML allow for flexibility in the RQs and methods if impassible constraints arise. Furthermore, with a pragmatic ideology, qualitative analysis would also be permissible, either in sequence, such as a quantitative big data analysis of project metrics followed by qualitative analysis of PM, sponsor and team member interviews (or surveys) to dig deeper into the phenomena.

Finally, recall from the literature review that approximately half or more of all IT-related projects around the world fail, for example the Standish Group reported only 32% were successful in 2009 (Standish-Group, 2009). The project success rates in China are even lower, at 22% (Hu, Zhang, Sun, Liu, & Du, 2009), resulting in a 78% failure rate. Given these high project failure rates combined with the lack of causal factor evidence, the authors recommend a think-out-side-the-box approach be considered to address this. With no other statistically significant causal factors being available, the authors assert that the high project failure rate could lie in part due to the lack of good quality project management discipline research and training. It was clear from the literature review search that good quality papers of project failure or success factors were almost entirely published outside of the peer-reviewed journals in the project management discipline. The PM-discipline-related journals cited in the current study with good quality empirical studies included: the *Journal of Engineering, Project, and Production Management*, the *International Journal of Information Technology Project Management*, the *International Journal of Project Organisation and Management*, and the *International Journal of Procurement Management*. These can be seen in the reference section. Proquest and EBSCO indexes were used, both provide excellent search engines free to the public including links to full text if available. These indexes should have discovered papers related

to the keywords discussed earlier. We found for example a relevant paper by PM practitioners Saadé et al. (2015) which was a well-designed empirical study published outside of PM industry journals, so that article and others like it have probably been overlooked by PM researchers.

Based on the above, the authors wonder if the project management association leaders and their journal managers do not believe the Standish analysis that roughly half or more of all projects around the world have been failing for more than five decades (Kurek, Johnson, & Mulder, 2017; Masticola, 2007; Standish-Group, 2009). The authors assert this lack of empirical attention by PM associations and journals towards studying IT-project failure short-changes the project management practitioners and researchers because it ignores valuable organizational data on project performance. The authors assert the project management profession needs to consider a scientific fact-finding data-driven philosophy. The authors argue that case studies and surveys are not accurate methods to investigate hypotheses of project failure in the current context when actual evidence likely exists in organizational databases. If the organization does not capture and record project performance metrics, then the argument is that organization is not likely suitable as a source of data for empirical investigation into the project failure factors. Such an organization may certainly be an ideal case study for other RQs but not specifically similar to the RQs in the current study. The authors also argue that single case studies are not generalizable beyond the site or company targeted. Remember the example of the megaproject case study about an oil platform where the implications can be generalized only to a comparable population – another offshore oil rig. The authors assert that at least comparative multiple case studies ought to be encouraged as a qualitative approach, but that should be balanced with more quantitative organizational data-driven analysis, or mix methods within a comparative multiple case study, in situations for studying IT-related project failure in the current state of the field. At the same time, we do not state that using single case studies as a method is bad – no – we assert single case studies are ideal especially for purposively selecting IT project best-practices at a high performing site (or vice-versa, a notable failure) with project success (or failure) embedded as an implied dependent variable.

Another thought was perhaps the project management-related higher education sector is also partly responsible for the high project failure rates. A salient point Karanja and Malone (2021) made was that the college curriculum for teaching project management needs to be dramatically improved. They noted that many of the learning outcomes in the syllabi they reviewed during 2016-2018 addressed lower levels of cognitive learning and the syllabi did not have well-written measurable outcomes. The professors teaching project management did not necessarily have project management industry experience and/or they were not often certified in any of the project management-related knowledge areas. The authors concur that better training must be given to project managers, to students in college as well as to employees in organizations. The authors assert the standards are too low in higher education for teaching project management. As suggested by Karanja and Malone (2021), the authors assert a university professor teaching project management ought to have relevant education/certification and hands-on experience in project management. As an analogy, if a professor were hired to teach brain surgery, would the institution want to select a certified/educated medical practitioner with at least some actual hands-on experience in brain surgery, or would it be good enough to have certification or experience with foot surgery? In closing the authors assert that more empirical peer-reviewed studies are needed to investigate project failure, and from other countries outside the U.S. where the current study took place.

CONFLICT OF INTEREST

The authors of this publication declare there is no conflict of interest.

FUNDING AGENCY

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Anthopoulos, L., Reddick, C. G., Giannakidou, I., & Mavridis, N. (2016). Why e-government projects fail? An analysis of the Healthcare.gov website. *Government Information Quarterly*, *33*(1), 161–173. doi:10.1016/j.giq.2015.07.003

Biba, M., Ballhysa, E., Vajjhala, N. R., & Mullagiri, V. R. (2010). *A novel structure refining algorithm for statistical-logical models.* Paper presented at the 2010 International Conference on Complex, Intelligent and Software Intensive Systems, Krakow, Poland.

Borbath, M., Blessner, P., & Olson, B. (2019). An empirical approach to the evaluation of defence contractor performance. *International Journal of Logistics Research and Applications*, *22*(2), 138–153. doi:10.1080/13675567.2018.1497148

Catanio, J. T., Armstrong, G., & Tucker, J. (2013). The effects of project management certification on the triple constraint. *International Journal of Information Technology Project Management*, *42*(1), 31–41. doi:10.4018/ijitpm.2013100106

Crosby, P. (2012). Characteristics and techniques of successful high-technology project managers. *International Journal of Project Organisation and Management*, *4*(2), 99–122. doi:10.1504/IJPOM.2012.046325

Eckerd, A., & Snider, K. (2017). Does the program manager matter? New public management and defense acquisition. *American Review of Public Administration*, *47*(1), 36–57. doi:10.1177/0275074015596376

Fatima, T., Azam, F., Anwar, M. W., & Rasheed, Y. (2020). *A systematic review on software project scheduling and task assignment approaches*. Paper presented at the Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence, Tianjin, China. doi:10.1145/3404555.3404588

Ghossein, T., Islam, A. M., & Saliola, F. (2018). Public procurement and the private business sector: Evidence from firm-level data. In K. V. Thai, P. J. Salia, & F. A. Mwakibinga (Eds.), *Global public procurement: Theories and practices* (pp. 21–55). Springer. doi:10.1596/1813-9450-8575

Goel, A. (2018). Analysing contractor prequalification criteria in construction: New insights through interpretive structural modelling. *International Journal of Procurement Management*, *11*(6), 667–683. doi:10.1504/IJPM.2018.095646

Han, W., Lung, C.-H., & Ajila, S. (2016). Using source code and process metrics for defect prediction - A case study of three algorithms and dimensionality reduction. *Journal of Software*, *11*(9), 883–902. doi:10.17706/jsw.11.9.883-902

Hu, Y., Zhang, X., Sun, X., Liu, M., & Du, J. (2009). An intelligent model for software project risk prediction. Paper presented at the *2009 International Conference on Information Management, Innovation Management and Industrial Engineering, Europe*, (pp. 629-632). IEEE. doi:10.1109/ICIII.2009.157

Huang, Z., & Cappel, J. J. (2018). A critical analysis of an entry level information systems certification. *Journal of Business and Educational Leadership*, *7*(1), 30–41.

Israel, J. W. (2012). Why the FBI can't build a case management system. *Computers*, *45*(6), 73–80. doi:10.1109/MC.2012.2

Jennings, W., Lodge, M., & Ryan, M. (2018). Comparing blunders in government. *European Journal of Political Research*, *57*(1), 238–258. doi:10.1111/1475-6765.12230

Karanja, E., & Malone, L. C. (2021). Improving project management curriculum by aligning course learning outcomes with Bloom's taxonomy framework. *Journal of International Education in Business*, *14*(2), 197–218. doi:10.1108/JIEB-05-2020-0038

Kurek, E., Johnson, J., & Mulder, H. (2017). Measuring the value of enterprise architecture on IT projects with CHAOS research. *Systems, Cybernetics, and Informatics*, *15*(7), 13–18.

Laurie, H. D., Rana, N. P., & Simintiras, A. C. (2017). The changing landscape of IS project failure: An examination of the key factors. *Journal of Enterprise Information Management*, *30*(1), 142–165. doi:10.1108/JEIM-01-2016-0029

Martinez-Perales, S., Ortiz-Marcos, I., Ruiz, J. J., & Lazaro, F. J. (2018). Using certification as a tool to develop sustainability in project management. *Sustainability*, *10*(5), 14–23. doi:10.3390/su10051408

Masticola, S. P. (2007). A simple estimate of the cost of software project failures and the breakeven effectiveness of project risk management. Paper presented at the *First International Workshop on the Economics of Software and Computation*. IEEE. doi:10.1109/ESC.2007.1

Memeti, S., Pllana, S., Binotto, A., Kołodziej, J., & Brandic, I. (2018). *A review of machine learning and meta-heuristic methods for scheduling parallel computing systems*. Paper presented at the Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications, Rabat, Morocco. doi:10.1145/3230905.3230906

Momoh, F. O., Rakshit, S., & Vajjhala, N. R. (2022, 2022//). *Exploratory study of machine learning algorithms in recommender systems.* Paper presented at the Proceedings of International Conference on Advanced Computing Applications, Singapore. doi:10.1007/978-981-16-5207-3_48

Nazeer, J., & Marnewick, C. (2018). Investing in project management certification: Do organisations get their moneys worth? *Information Technology Management*, *19*(1), 51–74. doi:10.1007/s10799-017-0275-y

Ngonda, V. S., & Jowah, L. E. (2020). A study of the impact of project managers' power and influence on their organisation's project management maturity, *MATEC Web of Conferences (Vol. 312).* South Africa: EDP Sciences. doi:10.1051/matecconf/202031203001

O'Boyle, E., Banks, G. C., Carter, K., Walter, S., & Yuan, Z. (2019). A 20-year review of outcome reporting bias in moderated multiple regression. *Journal of Business and Psychology*, *34*(1), 19–37. doi:10.1007/s10869-018-9539-8

Oliveira, A. L. (2019). Biotechnology, big data and artificial intelligence. *Biotechnology Journal*, *14*(8), 45–53. doi:10.1002/biot.201800613 PMID:30927505

Pace, M. (2019). A correlational study on project management methodology and project success. *Journal of Engineering, Project, and Production Management*, *9*(2), 56–65. doi:10.2478/jeppm-2019-0007

Patil, A. A., & Gogte, J. (2020). Importance ranking of supplier selection scorecard parameters - among Indian car manufacturing industry. *International Journal of Procurement Management*, *13*(4), 665–677. doi:10.1504/IJPM.2020.110081

Pospieszny, P., Czarnacka-Chrobot, B., & Kobylinski, A. (2018). An effective approach for software project effort and duration estimation with machine learning algorithms. *Journal of Systems and Software*, *137*(2), 184–196. doi:10.1016/j.jss.2017.11.066

Saadé, R. G., Dong, H., & Wan, H. (2015). Factors of project manager success. *Interdisciplinary Journal of Information, Knowledge, and Management*, *10*(1), 63–80. doi:10.28945/2265

Standish-Group. (2009). *Chaos Summary 2009: The 10 Laws of CHAOS*. University of Chicago. https://www.classes.cs.uchicago.edu/archive/2014/fall/51210-1/required.reading/Standish.Group.Chaos.2009.pdf

Strang, K. D. (2015). Matching research method with ideology and strategy. In K. D. Strang (Ed.), *Palgrave Handbook of Research Design in Business and Management* (pp. 47–62). Palgrave Macmillan. doi:10.1057/9781137484956_4

Strang, K. D. (2021). Which organizational and individual factors predict success versus failure in procurement projects. *International Journal of Information Technology Project Management*, *12*(3), 19–39. doi:10.4018/IJITPM.2021070102

Strang, K. D., & Perez, M. L. (2020). Statistical analysis of US government supply chain contract breaches. *Journal of Supply Chain Management. Logistics and Procurement*, *3*(3), 272–285.

Strang, K. D., & Sun, Z. (2022). Managerial controversies in AI and big data. In M. Khosrow-Pour (Ed.), Research Anthology on Big Data Analytics, Architectures, and Applications (pp. 1745-1764 [ch. 1785].PA, USA: Information Resource Management Association (IRMA). doi:1710.4018/1978-1741-6684-3662-1742.ch1085

Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., & Fischl, M. (2021). Artificial intelligence in supply chain management: A systematic literature review. *Journal of Business Research*, *122*, 502–517. doi:10.1016/j.jbusres.2020.09.009

USGAO. (2020). *Reports & testimonies. [Annual Report]*. United States Government Accountability Office (USGAO). https://www.gao.gov/reports-testimonies/

Wang, M., Fu, W., He, X., Hao, S., & Wu, X. (2020). A survey on large-scale machine learning. *IEEE Transactions on Knowledge and Data Engineering*, 1–12. doi:10.1109/TKDE.2020.3015777

Kenneth Strang is a globally recognized scholar who teaches and researches: business administration, economics-statistics, management, marketing-consumer behavior, human resource management, operations research, project management, organizational behavior, and supply chain management, plus he supervises doctoral students. Dr. Strang has over 37 years of career experience. He has numerous scholarly publications including a research methods best-selling textbook, analytics, learning styles/culture, leadership and he was a contributor to the ISO/ Project Management Body of Knowledge

Narasimha Rao Vajjhala is working as an Associate Professor and the Chair of the Information Systems department at the School of IT and Computing at the American University of Nigeria. He had previously worked at the Faculty of Engineering and Architecture at the University of New York Tirana, Albania. He is a senior member of ACM and IEEE. He is the Managing Editor for the International Journal of Risk and Contingency Management (IJRCM). He is also a member of the Risk Management Society (RIMS), and the Project Management Institute (PMI). He has over 20 years of experience in teaching mainly programming and database-related courses at both graduate and undergraduate levels in Europe and Africa. He has also worked as a consultant in technology firms in Europe and has experience participating in EU-funded projects. He has supervised several undergraduate senior design projects and graduate theses. He has completed a Doctorate in Information Systems and Technology (United States); holds a Master of Science in Computer Science and Applications (India), and a Master of Business Administration with a specialization in Information Systems (Switzerland), and an undergraduate degree in Computer Science (India).