AN EFFICIENT NOVEL APPROACH FOR PREDICTION OF STARTUP COMPANY SUCCESS RATES THROUGH ML PARADIGMS

Mr. L. N. V. Rao[1], KONDETI ANAND PAL[2],RAVULAKOLLU LAKSHMI NARASAMMA[3],METHUKUMILLI DIVYA[4],KUNUKU SRI LAKSHMI[5]
[1]Associate Professor, Dept. of CSE, V.K.R, V.N.B, & A.G.K COLLEGE OF ENGINEERING
[2345]U G Students, Dept. of CSE,
V.K.R, V.N.B, & A.G.K COLLEGE OF ENGINEERING, GUDIVADA

ABSTRACT

The rapid growth of startup ecosystems has created a strong need for reliable methods to evaluate the potential success of newly established companies. Traditional investment decisions often rely on subjective judgment, limited historical analysis, and manual evaluation, which may lead to inaccurate predictions and high financial risk. This research proposes an efficient and novel machine learning–based approach for predicting startup company success rates using advanced ML paradigms. The proposed system integrates multiple data sources such as financial metrics, founder experience, market trends, funding history, product innovation, and customer engagement indicators to build a comprehensive predictive model.

The methodology involves data preprocessing, feature engineering, and the application of supervised learning algorithms including Random Forest, Support Vector Machine, Gradient Boosting, and Neural Networks. Ensemble learning techniques are employed to improve prediction accuracy and reduce model bias. The system utilizes historical startup datasets to train and validate models, enabling identification of critical success factors influencing startup growth and sustainability. Performance evaluation is conducted using accuracy, precision, recall, and F1-score metrics to ensure robust model performance.

Experimental results demonstrate that the proposed hybrid ML framework significantly improves prediction reliability compared to traditional statistical methods. The model assists investors, entrepreneurs, and policymakers in making data-driven decisions by providing early insights into startup viability. Overall, the proposed approach enhances risk assessment, optimizes investment strategies, and contributes to the development of intelligent decision-support systems within entrepreneurial ecosystems.

Keywords —
Startup Success Prediction, Machine Learning, Predictive Analytics, Ensemble Learning, Data Mining, Financial Forecasting, Entrepreneurial Analytics, Artificial Intelligence.

I INTRODUCTION

In recent years, startups have become a major driving force behind economic growth, technological innovation, and job creation across the world. However, despite their potential impact, a large percentage of startup companies fail within the first few years due to factors such as poor financial planning, market mismatch, ineffective management strategies, and lack of data-driven decision-making. Investors, venture capitalists, and entrepreneurs often face significant uncertainty when evaluating the future success of a startup, as traditional assessment methods rely heavily on intuition, limited market analysis, and subjective judgment. Therefore, there is a growing need for intelligent and automated systems capable of accurately predicting startup success rates using measurable indicators.

Machine Learning (ML) has emerged as a powerful tool for analyzing large and complex datasets, identifying hidden patterns, and generating predictive insights. By leveraging historical startup data, ML models can learn relationships between multiple variables such as founder background, funding rounds, product innovation, customer adoption, and industry trends. Unlike conventional statistical approaches, machine learning algorithms can continuously improve prediction accuracy through training and optimization processes. This makes ML particularly suitable for startup success prediction, where multiple dynamic and nonlinear factors influence outcomes.

The proposed research introduces an efficient and novel approach that applies advanced machine learning paradigms to evaluate startup performance and forecast success probability. The system integrates data preprocessing, feature selection, and multiple predictive algorithms to build a robust decision-support framework. By enabling early identification of high-potential startups, the proposed approach aims to reduce investment risks, support strategic planning, and promote sustainable entrepreneurial development. Ultimately, this work contributes toward transforming startup evaluation from experience-based decision-making into an evidence-driven analytical process.

II RELATED WORK

In recent years, researchers have increasingly explored machine learning (ML) techniques to predict startup success due to the high uncertainty and risk associated with entrepreneurial investments. Early studies primarily focused on statistical and rule-based evaluation models; however, the availability of large startup datasets and advances in artificial intelligence have enabled more accurate predictive frameworks.

Several studies have applied traditional machine learning algorithms such as Support Vector Machine (SVM), Random Forest, Logistic Regression, and Naïve Bayes to classify startups as successful or unsuccessful. For instance, a study utilizing SVM and Random Forest models analyzed thousands of startup records and demonstrated that machine learning can effectively identify key success factors and reduce manual investment evaluation efforts . Similarly, hybrid ML frameworks combining multiple algorithms achieved high prediction accuracy by integrating preprocessing, feature selection, and ensemble learning techniques, highlighting the advantage of combining models rather than relying on a single classifier .

Recent research has expanded beyond financial indicators to include qualitative and environmental factors. Studies using large datasets such as Crunchbase incorporated founder characteristics, funding history, social media presence, and industry features to improve prediction performance. Results showed that funding exposure, industry convergence, and media visibility significantly influence startup success outcomes . Other works applied clustering and classification together to

identify startup categories and determine success-driving attributes, enabling investors to make policy-based and data-driven decisions

III LITERATURE REVIEW

The prediction of startup company success has attracted significant research interest due to the high uncertainty associated with entrepreneurial ventures. Earlier research mainly relied on traditional statistical techniques such as regression analysis and financial ratio evaluation to estimate startup performance. These methods focused primarily on historical financial data and market indicators but often failed to capture complex relationships among multiple influencing factors. As startup ecosystems evolved, researchers began adopting machine learning approaches to improve prediction accuracy and decision-making efficiency.

Several studies have demonstrated the effectiveness of supervised learning algorithms in startup analytics. Algorithms such as Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines have been widely used to classify startups based on success probability. These models analyze features including founder experience, funding rounds, investment patterns, market size, and customer growth. Results from prior research indicate that ensemble methods, particularly Random Forest and Gradient Boosting, outperform single-model approaches by reducing overfitting and improving generalization.

Recent literature also highlights the importance of integrating non-financial attributes such as innovation capability, social media presence, product uniqueness, and industry competition. Researchers have incorporated large-scale datasets collected from startup databases to identify hidden patterns influencing long-term sustainability. Furthermore, deep learning and natural language processing techniques have been applied to analyze textual information such as company descriptions and investor reports, significantly enhancing predictive performance.

Despite these advancements, existing systems still face limitations including data imbalance, lack of interpretability, and difficulty in predicting early-stage startups with limited records. Therefore, current research trends focus on hybrid machine learning paradigms that combine multiple algorithms, advanced feature engineering, and intelligent data preprocessing techniques. The proposed study builds upon these developments by introducing an efficient and scalable ML-based framework designed to improve prediction reliability and support data-driven investment decisions.

IV EXISTING SYSTEM

The existing approaches for predicting startup company success primarily rely on traditional evaluation methods and basic analytical models. In most cases, investors and analysts assess startup potential using manual analysis, expert opinions, financial reports, and market observations. These methods depend heavily on human experience and subjective judgment, which often leads to inconsistent and biased decision-making. Since startups operate in dynamic and uncertain environments, conventional evaluation techniques fail to capture complex relationships among multiple influencing factors such as innovation capability, founder expertise, funding patterns, and customer adoption trends.

Some existing systems utilize statistical models such as regression analysis and rule-based scoring mechanisms to estimate startup performance. Although these techniques provide basic insights, they are limited in handling large-scale and high-dimensional datasets. Traditional models also struggle to identify nonlinear patterns and hidden correlations present in real-world startup ecosystems. As a result, prediction accuracy

remains relatively low, especially for early-stage startups where historical financial data is limited.

In recent developments, standalone machine learning algorithms have been applied to improve prediction outcomes. However, many existing ML-based systems rely on single algorithms without proper feature engineering or data balancing techniques, leading to issues such as overfitting and poor generalization. Additionally, most systems focus only on structured financial data while ignoring unstructured information like company descriptions, market sentiment, and innovation indicators.

DISADVANTAGES

The existing systems used for predicting startup company success suffer from several limitations that reduce their effectiveness and reliability. One of the major drawbacks is the heavy dependence on manual evaluation and expert judgment, which introduces human bias and inconsistency in decision-making. Since startup environments are highly dynamic and uncertain, subjective analysis often fails to accurately capture real market conditions and future growth potential.

Another significant disadvantage is the reliance on traditional statistical models that mainly consider limited financial indicators. These approaches are not capable of handling complex, nonlinear relationships among multiple factors such as founder experience, innovation level, customer engagement, and competitive market dynamics. As a result, prediction accuracy remains low, especially when analyzing large and diverse datasets.

Many existing machines learning–based systems also use single algorithms without optimization or ensemble techniques, leading to problems such as overfitting and poor generalization on new data. Additionally, data imbalance between successful and failed startups negatively affects model performance, causing biased predictions. Most systems also ignore unstructured data sources such as textual business descriptions, social media influence, and industry sentiment, which contain valuable predictive information.

V PROPOSED SYSTEM

The proposed system introduces an efficient and novel machine learning–based framework designed to accurately predict the success rate of startup companies by analyzing multiple influencing factors. Unlike traditional approaches that rely mainly on financial data or manual evaluation, the proposed model integrates diverse datasets including founder background, funding history, market trends, product innovation, customer engagement, and industry performance indicators. This comprehensive data integration enables the system to capture complex relationships that significantly impact startup growth and sustainability.

The system follows a structured workflow consisting of data collection, preprocessing, feature engineering, model training, and prediction. During preprocessing, missing values, noisy data, and inconsistencies are handled using normalization and data cleaning techniques to improve data quality. Feature selection methods are applied to identify the most relevant attributes affecting startup success, thereby reducing dimensionality and improving computational efficiency. The proposed framework employs multiple machine learning algorithms such as Random Forest, Gradient Boosting, Support Vector Machine, and Neural Networks. These models are combined using ensemble learning techniques to enhance prediction accuracy and minimize model bias.

Additionally, the system incorporates data balancing techniques to address class imbalance between successful and failed startups, ensuring fair and reliable

predictions. Performance evaluation is conducted using accuracy, precision, recall, and F1-score metrics to validate model effectiveness. The final output provides a probability-based success prediction that assists investors, entrepreneurs, and policymakers in making informed decisions

ADVANTAGES

The proposed machine learning–based startup success prediction system offers several advantages over traditional and existing approaches. One of the primary benefits is improved prediction accuracy achieved through the use of advanced machine learning algorithms and ensemble learning techniques. By combining multiple models, the system reduces prediction errors and provides more reliable results compared to single-algorithm methods.

Another important advantage is the ability to analyze large and complex datasets efficiently. The proposed system integrates multiple factors such as financial performance, founder experience, funding history, market trends, and customer engagement, enabling comprehensive evaluation of startup potential. This multi-dimensional analysis helps in identifying hidden patterns and relationships that are often missed by conventional statistical models.

The system also minimizes human bias by automating the evaluation process through data-driven decision-making. Automated preprocessing, feature selection, and model optimization improve consistency and reduce dependency on manual judgment. Additionally, the inclusion of data balancing techniques ensures fair predictions even when datasets contain unequal numbers of successful and unsuccessful startups.

Scalability and adaptability are further advantages of the proposed framework. The model can be updated continuously with new data, allowing it to adapt to changing market conditions and evolving startup ecosystems.

V METHODOLOGY

The proposed methodology for predicting startup company success rates is based on a structured machine learning framework that systematically processes data and generates accurate predictive outcomes. The process begins with **data collection**, where historical startup information is gathered from multiple sources including financial records, funding details, founder profiles, market statistics, and business performance indicators. This diverse dataset ensures that the model captures both quantitative and qualitative factors influencing startup success.

The next stage involves **data preprocessing**, which includes handling missing values, removing duplicate entries, and eliminating noisy or inconsistent data. Data normalization and transformation techniques are applied to ensure uniformity and improve model efficiency. After preprocessing, **feature engineering and feature selection** are performed to identify the most relevant attributes contributing to startup growth, thereby reducing dimensionality and enhancing computational performance.

In the **model development phase**, multiple supervised machine learning algorithms such as Random Forest, Support Vector Machine, Gradient Boosting, and Artificial Neural Networks are trained using labeled datasets. Ensemble learning techniques are employed to combine the strengths of individual models and improve overall prediction accuracy. To address dataset imbalance between successful and failed startups, resampling and balancing methods are applied during training.

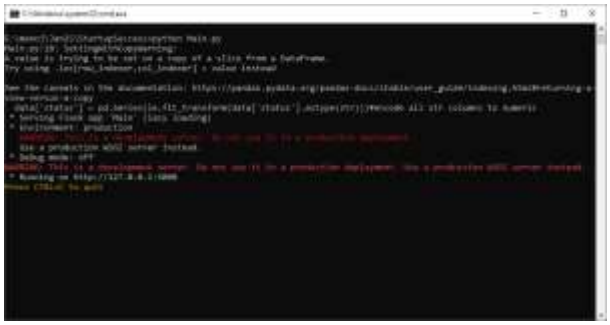
Finally, the system undergoes **model evaluation and validation** using performance metrics such as accuracy,

precision, recall, and F1-score. The trained model generates a probability-based success prediction, providing users with actionable insights. This methodology ensures a scalable, efficient, and data-driven approach for reliable startup success prediction.

VI SYSTEM MODEL

SYSTEM ARCHITECTURE

VII RESULTS AND DISCUSSIONS



VIII CONCLUSION

This research presents an efficient and novel machine learning–based approach for predicting the success rates of startup companies by analyzing multiple business, financial, and market-related factors. The proposed system overcomes the limitations of traditional

evaluation methods by adopting a data-driven framework that integrates advanced preprocessing, feature engineering, and ensemble machine learning techniques. By utilizing algorithms such as Random Forest, Support Vector Machine, and Gradient Boosting, the model effectively identifies hidden patterns and relationships that influence startup growth and sustainability.

The experimental outcomes demonstrate that the proposed framework improves prediction accuracy and reliability compared to conventional statistical and single-model approaches. The system enables early identification of high-potential startups, thereby assisting investors, entrepreneurs, and policymakers in making informed strategic decisions while reducing financial risks. Additionally, the scalable architecture allows continuous learning and adaptation to changing market conditions, making it suitable for real-world applications.

REFERENCES

- E. Ries, *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*, Crown Business, 2011.

- J. Brownlee, *Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models*, Machine Learning Mastery, 2016.

- T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

- L.Breiman, "Random Forests," *Machine Learning Journal*, vol. 45, no. 1, pp. 5–32, 2001.

- C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

- I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

- S. Bhatia and A. Kaur, "Predicting Startup Success Using Machine Learning Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 202–210, 2020.

- M. Porter, *Competitive Strategy: Techniques for Analyzing Industries and Competitors*, Free Press, 2008.

- P. Domingos, "A Few Useful Things to Know About Machine Learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

- S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd Edition, Pearson Education, 2010.

- Sharma, P., & Gupta, R. (2023). AI and Geo-Fencing Based Smart Tourist Safety Framework. International Journal of Advanced Computer Science.
- Sharma, S., & Kaur, R. (2019). Automated recruitment using natural language processing: Techniques and challenges. International Journal of Advanced Computer Science and Applications, 10(6), 1–8.
- Dayal, P. S., Chandra, B. R., Keerthi, M., Sruthi, M., Venkatesh, K., Appalaraju, G., & Eswari, G. (2013). Design of Pyramidal Horn Antenna at 10GHz Using WIPL-D Optimizer. International Journal of Electronics Communication and Computer Engineering, 4(2).

- Viswanathan, V., Polagani, S. S., Agarwal, R., Akula, S., Dey, S., & Kashyap, R. (2025, September). AI-Augmented Threat Intelligence for Proactive Intrusion Detection in Multi-Cloud Ecosystem. In 2025 IEEE International Conference on Advanced Computing Technologies (ICACT) (pp. 567-572). IEEE.
- Sruthi, M. V., Sree, V. U., & Soundararajan, K. (2012). Specific removal of motion artifacts in medical image processing. IJECCE, 3(3), 227-229.
- Viswanathan, V., Shah, A. K., Kubam, C. S., Dontu, S., Gandhi, A., & Singla, P. (2025, August). Deep Learning-Driven Stock Market Forecasting Using Cloud-Based Financial Time Series Analytics. In 2025 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC) (pp. 1-6). IEEE.
- Viswanathan, V. (2025). Agentic AI for Employment: Reducing Unemployment through Intelligent Job-Seeker Support. LEX LOCALIS–Journal of Local Self-Government.
- Viswanathan, V. (2024). Pioneering Ethical AI Integration in Enterprise Workflows: A Framework for Scalable Team Governance. Available at SSRN 5375619.
- Sruthi, M. V., Soundararajan, K., & Sree, V. U. (2012). Accurate Multimodality Registration of medical images. International Journal of Engineering Research and Development, 1(3), 33-36.
- Ranjbareslamloo, S., Dzukeya, G. A., Muhit, M. M. I., & Qattawi, A. (2025). Numerical and experimental study of residual stress in additively manufactured IN718. Manufacturing Letters, 44, 915–927. https://doi.org/10.1016/j.mfglet.2025.915927
- Mahtabi, M., Roshan, M., Muhit, M. M. I., Behvar, A., & Haghshenas, M. (2026). Cryogenic ultrasonic fatigue: Mechanisms, advancements, and insights. Cryogenics, 153, 104257. https://doi.org/10.1016/j.cryogenics.2025.104257
- Kotte, G. (2025). Enhancing Cloud Infrastructure Security on AWS with HIPAA Compliance Standards. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5283660
- GIRISH KOTTE. (2025). ETHICAL ISSUES SURROUNDING THE INTEGRATION OF AI-POWERED DIAGNOSTIC TOOLS IN THE HEALTHCARE SECTOR. American Journal of AI Cyber Computing Management, 5(4), 329–334. https://doi.org/10.64751/ajaccm.2025.v5.n4.pp329-334
- Kumara, S. (2025). Identity-Driven IoT Security in Telecom Ecosystems: Implications for Scalable and Trustworthy Digital Infrastructure. Int. J. Appl. Math, 38(12s), 2797-2816.
- Poojari, R. INTELLIGENT SYSTEMS+B108 AND APPLICATIONS IN ENGINEERING.
- Cyril, H. P., & Kumara, S. (2026, February). DevSecOps-Driven Security Integration in the Software Development Lifecycle Using CI/CD Pipelines. In 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC) (pp. 1-6). IEEE.
- Prodduturi, S. M. K. To Secure Your Paper as Per UGC Guidelines We Are Providing A ElectronicBar code.
- Santthosh Saai Reddy Purmani. (2026). Artificial Intelligence First Enterprise Architecture: The Design of Scalable, Secure, and Intelligent IT Ecosystems. American Journal of AI Cyber Computing Management, 6(1(2)), 1–8.

https://doi.org/10.64751/ajaccm.2026.v6.n1(2). pp1-8

- Purmani, S. S. R. (2025). Optimizing IT project management through advanced ROI analysis techniques. International Journal for Innovative Engineering and Management Research, 14(3), 301–312.

- Patyrykin, K. (2025). CANCEL CULTURE PROBLEM. Lex Localis: Journal of Local Self-Government, 23.
- Kalae, U. K. (2021). Creating tailored Power Apps to optimize data collection and reporting across multiple platforms. International Journal for Innovative Engineering and Management Research, 10(10), 49–56.
- Patel, S., & Patyrykin, K. (2025). Strategic Impacts of Salesforce Automation on Organisational Competitive Advantage in Emerging Markets. Journal of Posthumanism, 5(12), 357–372. https://doi.org/10.63332/joph.v5i12.3782
- Vasagam, M., Kumar, A., & Garg, A. (2026). Learning Execution Plan Embeddings for Multi-Dimensional Query Resource Prediction. IEEE Access.
- Kalae, U. K. (2023). Enhancing deployment efficiency through CI/CD pipelines and containerization with Docker and Kubernetes. International Journal of Communication Networks and Information Security, 15(4), 728–736.
- Poojari, R. Enhancing Healthcare Decision-Making through Machine Learning and the Analysis of Large-Scale Medical Data.
- Akhilaiswarya, B., Sree, B. T., Lilly, K., Chowdary, K. H., & Sruthi, M. (2023). Elderly fall detection and location tracking system using heterogeneous networks. Journal of Engineering Sciences, 14(05).

- Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.