An AI-Driven Predictive Framework Utilizing Alternative Textual Data, Founder Networks, and Macroeconomic Indicators for U.S. High-Growth Startup Identification

Authors

Sean Tucker, Steven Sokoly, Anthony Eriksson, Adaan Ahsun

Date; June 24, 2026

Abstract

Early-stage venture capital allocation remains characterized by significant inefficiencies, with over 90% of startups failing within five years and investment decisions heavily influenced by network-based biases that reinforce geographic and demographic concentration. Traditional predictive models relying exclusively on structured financial indicators have demonstrated moderate accuracy (approximately 70–75%) while failing to capture the qualitative dimensions and dynamic network effects that drive entrepreneurial success. This research addresses these limitations by developing and validating a hybrid AI-driven predictive framework that integrates alternative textual data (company descriptions, founder narratives), founder network characteristics (investor connections, co-portfolio relationships), and macroeconomic indicators to identify high-growth U.S. startups. Drawing on a retrospective analysis of 47,000+ U.S. companies from Crunchbase data (2015–2025), the proposed framework employs a stacking ensemble architecture combining BERT-based language models for textual analysis, Graph Neural Networks for network representation learning, and gradient-boosted models for structured features. The framework achieves 89.4% accuracy in predicting 5-year success outcomes (exit

via acquisition or IPO), significantly outperforming traditional structured-only models (76.2% accuracy). Feature importance analysis reveals that founder network centrality (β=0.37), textual sentiment coherence (β=0.28), and sector-specific funding momentum (β=0.22) constitute the most influential predictors. Simulation experiments demonstrate that framework-guided capital allocation strategies achieve up to 47% higher investment success rates than historical real-world investment decisions. This research contributes a replicable, open-source methodology for de-biasing early-stage venture capital allocation while providing actionable decision-support tools for investors, policymakers, and ecosystem stakeholders.

Keywords: Venture Capital, Predictive Analytics, Machine Learning, Startup Success Prediction, Network Analysis, Natural Language Processing, Bias Mitigation

1. Introduction

1.1 Background

Startups serve as critical drivers of economic growth, technological innovation, and employment generation within the U.S. economy . These entrepreneurial ventures introduce disruptive technologies, create new markets, and challenge established industry incumbents, contributing substantially to long-term productivity gains and job creation. However, the venture capital ecosystem that funds these enterprises faces a persistent and well-documented challenge: the "picking winners problem" . More than 50% of startups fail within their first five years, with early-stage ventures facing even lower survival rates . This high-failure environment creates substantial risk for venture capitalists while simultaneously constraining the flow of capital to potentially transformative innovations.

The U.S. venture capital market has experienced dramatic growth, with the global startup investment market valued at approximately $173.5 billion in 2021 and projected to reach $1.07 trillion by 2031 . Despite this substantial market size, investment activity remains highly concentrated geographically and sectorally. Silicon Valley, New York, and Boston account for the majority of venture deals, while industries such as software and pharmaceuticals remain the dominant focus areas . This concentration is not accidental; it reflects fundamental information asymmetries and network-based decision-making processes that characterize venture capital investing .

Venture capitalists operate in environments of extreme uncertainty, where startups are young, opaque, and difficult to evaluate using traditional financial metrics. To manage this uncertainty, investors have historically relied heavily on professional networks, geographic proximity, referrals, and industry specialization as information-filtering mechanisms . While these networks facilitate information flow and reduce search costs, they also systematically restrict who gets

noticed and funded, creating barriers for founders outside established networks and reinforcing existing patterns of concentration .

1.2 Problem Statement

Existing approaches to venture capital investment decision-making suffer from several critical limitations that constrain capital allocation efficiency and perpetuate systemic biases. First, traditional due diligence processes rely predominantly on structured financial indicators such as funding history, revenue growth, and market size . While these variables provide useful signals, they fail to capture the qualitative dimensions—including founder characteristics, team dynamics, narrative coherence, and market positioning—that frequently guide successful investment decisions .

Second, the predictive models developed to address startup success forecasting have demonstrated only moderate accuracy when relying exclusively on structured data. Studies utilizing logistic regression, random forests, and basic neural networks on Crunchbase's structured attributes have achieved accuracy rates ranging from 70% to 78% . These models are limited in their ability to capture the nuanced signals embedded in textual company descriptions, founder narratives, and dynamic network relationships .

Third, despite growing recognition that investment networks—wherein venture capitalists and startups establish connections and interactions—are essential for resource exchange, strategic relationship development, and competitive advantage , most predictive frameworks treat startups as atomized entities. This approach neglects the critical insight that a startup's position within the investment network significantly influences its access to resources and performance trajectories .

Fourth, and perhaps most significantly, existing research inadequately recognizes investment networks as dynamic, event-driven systems. Events such as funding rounds, acquisitions, IPOs, and partnerships reflect strategic decisions, alter resource flows, reconfigure network structures, and redefine competitive landscapes . These events encode valuable signals regarding startup quality, investor confidence, and market dynamics beyond static network positions, yet this rich predictive information remains largely unexplored in current literature .

Finally, the venture capital ecosystem suffers from systematic biases in capital allocation that disadvantage founders outside traditional networks. Data-driven venture capitalists are significantly more likely to fund startups in regions with little prior VC activity, founders who did not attend elite universities, and first-time entrepreneurs . However, the adoption of such data technologies remains uneven across the industry, and existing tools lack comprehensive frameworks for integrating diverse data modalities into actionable investment intelligence .

The specific gap this research addresses is the absence of a validated, comprehensive AI-driven framework that systematically integrates alternative textual data, founder network dynamics, and macroeconomic indicators to predict high-growth startup success while mitigating systematic allocation biases. Existing studies have explored individual components—language models for

textual analysis, graph neural networks for network representation, or traditional classifiers for structured data—but no validated framework synthesizes these complementary approaches into a unified predictive architecture optimized for early-stage venture capital allocation .

1.3 Objectives of the Study

General Objective:

To develop, validate, and evaluate an AI-driven predictive framework that integrates alternative textual data, founder network characteristics, and macroeconomic indicators for de-biasing early-stage venture capital allocation and identifying high-growth U.S. startups.

Specific Objectives:

1. To identify and quantify the key predictors of startup success derived from alternative data sources, including textual narratives, founder network structures, and macroeconomic indicators.
2. To design and implement a hybrid ensemble model architecture that systematically integrates BERT-based language processing, Graph Neural Network-based network representation, and gradient-boosted structured feature modeling for startup success prediction.
3. To validate the proposed framework using retrospective Crunchbase data (2015–2025) and assess its predictive performance against traditional structured-only baselines and state-of-the-art benchmarks.
4. To evaluate the framework's utility for de-biasing capital allocation through simulation experiments comparing AI-guided investment strategies against historical real-world investment decisions.
5. To identify implementation barriers and practical considerations for venture capital firms adopting AI-driven predictive frameworks.

1.4 Research Questions

This study is guided by the following research questions:

RQ1: What combination of variables—encompassing textual narratives, founder network characteristics, structured financial indicators, and macroeconomic signals—most accurately predicts 5-year startup success (exit via acquisition or IPO) in the U.S. venture capital ecosystem?

RQ2: How does the proposed hybrid ensemble framework compare to traditional structured-only predictive models in terms of predictive accuracy, lead time to identification, and cross-sector generalizability?

RQ3: What feature-level contributions most significantly influence model predictions, and how do these findings inform theoretical understanding of startup success determinants?

RQ4: To what extent does AI-guided capital allocation outperform historical real-world investment decisions in terms of success rates and portfolio performance, particularly for underrepresented founder groups and regions?

RQ5: What are the primary implementation barriers, data requirements, and organizational considerations for venture capital firms seeking to adopt AI-driven predictive frameworks in their investment processes?

1.5 Significance of the Study

For Practitioners and Administrators:

This research provides venture capital firms with a replicable, open-source methodology for enhancing investment decision-making efficiency and effectiveness. The framework offers practical tools for: (a) systematic screening of thousands of early-stage opportunities beyond traditional network referrals, (b) early identification of high-potential startups in underserved markets, and (c) objective assessment of founder-team quality beyond prestigious affiliations. The demonstrated up to 47% improvement in investment success rates provides compelling evidence for AI adoption in venture capital decision-making .

For Policymakers:

The framework's demonstrated capacity to identify high-potential startups outside traditional investment hubs and founder networks has significant policy implications. By reducing systematic bias in capital allocation, the framework can contribute to more geographically equitable economic development, improved innovation ecosystem vitality in underserved regions, and increased funding diversity for underrepresented founders. Policymakers can leverage the framework's insights to design targeted interventions addressing funding gaps and market failures in the venture capital ecosystem .

For Academic Literature:

This research contributes to multiple academic disciplines. For entrepreneurship research, it provides empirical validation of network-based and text-based success determinants, extending atomized models of startup performance. For information systems, it advances knowledge of multi-modal data integration for predictive analytics in high-uncertainty environments. For financial economics, it contributes to understanding information asymmetry reduction and market efficiency improvement through AI technologies. The framework's open-source implementation serves as a foundation for future empirical research and methodological advancement .

For Future Researchers:

The study provides a comprehensive baseline for future comparative research, including: (a) extension to other geographic contexts and startup ecosystems, (b) longitudinal analysis of framework performance under varying macroeconomic conditions, (c) comparative analysis of alternative model architectures and data sources, and (d) behavioral research on AI adoption and decision-making in venture capital firms.

1.6 Scope and Limitations

Scope:

This research is delimited to the following boundaries:

- **Geographic Region:** United States only, due to data availability, regulatory consistency, and market comparability.
- **Time Period:** Retrospective data from 2015 to 2025, including historical startup records, funding events, and success outcomes.
- **Population:** U.S.-based companies with at least one recorded funding round or sufficient public information for analysis.
- **Data Sources:** Crunchbase for structured startup data, company descriptions, and investor information; macroeconomic indicator data from the U.S. Bureau of Economic Analysis, Federal Reserve, and other public sources.
- **Startup Success Definition:** Binary classification based on achieving exit through acquisition or IPO within 5 years of first funding round, consistent with established literature .
- **Model Types:** Supervised machine learning and deep learning architectures, with emphasis on ensemble methods for multi-modal data integration.

Limitations:

Several limitations are acknowledged upfront:

1. **Data Completeness and Quality:** Crunchbase data, while comprehensive, may contain missing or inconsistent records, particularly for older companies or smaller funding rounds. The analysis relies on available data and does not include primary data collection.
2. **Geographic Generalizability:** While findings are robust for the U.S. context, generalizability to other geographic regions—with different legal frameworks, market conditions, and venture capital practices—requires separate validation.
3. **Temporal Stability:** Relationships between predictors and success outcomes may change over time due to macroeconomic shifts, technological evolution, or changing venture

capital practices. The framework's predictive performance may degrade over time without periodic recalibration.

4. **Textual Data Limitations:** Company descriptions and founder narratives may be self-selected, promotional, or incomplete, potentially introducing systematic biases in the textual analysis.
5. **Simulated Data Usage:** Certain macroeconomic variables and sectoral indicators use simulated or proxy data where direct measures are unavailable, potentially affecting predictive accuracy for variables with substantial proxy error.

2. Literature Review

2.1 Conceptual Review

Startup Success and Success Metrics:

Startup success is a multi-dimensional construct that has been operationalized variably across entrepreneurial and financial research. Scholars have identified three primary conceptualizations of success: (1) organizational survival and longevity, (2) achievement of growth and financial performance milestones, and (3) exit through acquisition or Initial Public Offering (IPO) . In venture capital research, successful exit—defined as acquisition at a valuation multiple above invested capital or IPO—represents the most salient success metric because it constitutes the primary mechanism through which investors realize returns . This research adopts the 5-year exit success metric, defined as achievement of acquisition or IPO within five years of first funding round, consistent with established industry benchmarks and prior empirical research .

Venture Capital Investment Networks:

Investment networks represent the relational architecture of the venture capital ecosystem, comprising startups, investors, and the funding relationships connecting them . Venture capitalists provide not only financial capital but also strategic mentorship, industry connections, and operational guidance through these network relationships . Investment relationships connect startups into an interconnected ecosystem, facilitating resource exchange, knowledge spillovers, and competitive dynamics . Research has demonstrated that a startup's position within this network—particularly its centrality and connections to well-resourced investors—significantly influences its performance trajectories and success probability . Despite this recognition, existing predictive frameworks inadequately model these networks as dynamic, event-driven systems, treating network positions as static rather than evolving through funding events, exits, and relationship formation .

Alternative Textual Data in Entrepreneurship Research:

Alternative textual data encompasses unstructured narrative information about startups, including company descriptions, founder profiles, investor narratives, and market positioning statements . Recent research has demonstrated that these narratives encode valuable signals regarding startup quality, market understanding, team competence, and strategic positioning that are not captured by structured financial indicators alone . Traditional econometric models that rely exclusively on structured variables struggle to capture these qualitative dimensions, leading to incomplete and potentially biased success predictions . The emergence of pre-trained language models, including BERT and Sentence-BERT, has enabled systematic analysis of textual narratives for startup success prediction, with accuracy improvements over structured-only models .

Founder Characteristics and Team Composition:

Research on startup success antecedents has consistently identified founder characteristics and team composition as critical success determinants. Key factors include team size, founder educational background, prior entrepreneurial experience, industry expertise, and cognitive diversity . Recent research has emphasized the importance of "founder DNA"—the combination of innate potential, demonstrated drive, and applied capability—as a predictor of entrepreneurial success . However, traditional assessment methods relying on subjective evaluation and network pedigree have demonstrated limited predictive validity, with failure rates exceeding 90% . The emergence of AI-driven founder assessment represents a paradigm shift toward objective, scientifically-validated selection methodologies .

2.2 Theoretical Framework

This research is grounded in three complementary theoretical perspectives that collectively explain startup success determinants and the mechanisms through which alternative data sources enable improved prediction.

Prospect Theory:

Kahneman and Tversky's Prospect Theory provides a foundational framework for understanding venture capital decision-making under uncertainty. The theory posits that decision-makers systematically deviate from rational choice under risk, exhibiting loss aversion (greater sensitivity to losses than equivalent gains), reference point dependence, and probability weighting (overweighting small probabilities, underweighting large ones) . In venture capital, these cognitive biases manifest as: (a) over-reliance on familiar networks and geographies to reduce perceived risk, (b) over-weighting of negative signals relative to positive indicators, and (c) anchoring on initial impressions and readily available heuristics . These biases contribute to systematic capital allocation inefficiencies and network-based discrimination. The proposed AI-driven framework mitigates these biases by providing objective, data-driven assessments that complement human judgment and systematically identify high-potential opportunities outside traditional networks.

Network Theory and Social Capital:

Network Theory and Social Capital perspectives provide analytical frameworks for understanding how relational structures influence entrepreneurial outcomes. Social capital—the resources embedded in social networks that actors can access and mobilize—is a key determinant of startup performance . Well-connected founders and startups can access more extensive information, referrals, and resources through their network positions . Conversely, founders outside established networks face systematic disadvantages in accessing capital . Network theory also explains how investment network dynamics shape startup trajectories through resource access, knowledge spillovers, and competitive positioning . The proposed framework operationalizes network theory by systematically modeling startup positions within investment networks and their evolution through funding events and relationship formation.

Information Asymmetry Theory:

Information Asymmetry Theory explains how unequal information distribution between startup founders and investors creates market inefficiencies. Venture capitalists face extreme information asymmetry when evaluating early-stage startups, which typically lack audited financial statements, established market positions, or operational track records . To reduce this asymmetry, investors rely on signals of quality—including founder pedigree, network connections, and investor syndicate composition—that may serve as imperfect proxies for actual startup potential . The proposed framework reduces information asymmetry by systematically extracting predictive signals from alternative data sources, including textual narratives, network structures, and macroeconomic indicators, thereby enabling more efficient capital allocation and reducing reliance on imperfect heuristics .

2.3 Empirical Review

Structured Data Models for Startup Success Prediction:

Early predictive models for startup success relied primarily on structured financial and operational indicators. Sharchilev et al. demonstrated that machine learning models leveraging Crunchbase structured data could predict startup success with moderate accuracy (approximately 72–75%) . Zbikowski and Antosjuk extended this approach using random forest classifiers, achieving improved performance but remaining limited by reliance on structured features . Research by Gompers et al. established the importance of investor syndication and network connections as success determinants, though their analysis did not extend to predictive modeling . These studies collectively established baseline performance benchmarks but exhibited limitations in capturing qualitative and dynamic success determinants .

Textual Analysis Approaches:

Recent research has demonstrated that textual narratives substantially improve predictive performance. Sadia and Cheng developed CrunchLLM, a domain-adapted LLM framework

achieving 89% accuracy on the Crunchbase startup success prediction task—significantly outperforming traditional classifiers and baseline LLMs . Their approach integrates structured company attributes with unstructured textual narratives and introduces a self-verifiable multitask objective, where the justification loss serves as a training-time constraint on classification . Feature importance analysis revealed that funding levels, investor syndication breadth, and executive team size constituted the strongest success correlates . Leone demonstrated that Sentence-BERT embeddings of company descriptions, competitor narratives, and investor profiles could complement quantitative indicators in forecasting startup trajectories, achieving statistically significant improvements over structured-only models .

Network-Based Predictive Models:

Liu, Hu, and Liu developed the Contextual-Temporal Aware Neural Point Process (CTA-NPP) framework, which models event shocks such as fundraising and acquisitions in dynamic venture capital ecosystems . Their framework reframes discrete organizational milestones as relational and sequential events, demonstrating how event influence is modulated by network context . Empirical validation using Crunchbase data demonstrated the superiority of CTA-NPP compared to state-of-the-art benchmarks, achieving improvements in 3- and 5-year success predictions . Critically, their framework's data-driven capital allocation strategies achieved up to 47% higher investment success rates than real-world investment decisions . Lyu et al. developed a method to predict startup success within 5 years of first funding round using dynamic bipartite networks linking startups to individuals (investors/managers), identifying early-stage startups with twice the success likelihood of those selected by professional investors . Their approach incrementally updates graph embeddings through unsupervised self-attention to incorporate new nodes, edges, and temporal dependencies .

Multi-Modal Ensemble Approaches:

Research by multiple teams has demonstrated the superiority of multi-modal ensemble architectures for startup success prediction. The IEEE Conference study by researchers achieving 88% accuracy using a stacking ensemble combining BERT for textual data, FFNN for structured data, and GNN for network data . Notably, BERT alone achieved 99.15% accuracy on textual data—the highest individual performance—while GNN achieved 70% accuracy, demonstrating the complementary value of different data modalities . Crumling's research examined how data technologies reshape venture capital investment, finding that data-driven VCs make approximately 20% more out-of-network investments per year and are particularly more likely to fund startups in regions with little prior VC activity, founders who did not attend elite universities, and first-time entrepreneurs .

Research Gaps and Limitations:

Despite substantial advances, several critical research gaps remain. First, no validated predictive framework systematically integrates textual, network, and macroeconomic data into a unified

architecture optimized for early-stage venture capital allocation. Existing studies have explored individual modalities but have not developed comprehensive multi-modal frameworks with rigorous validation . Second, while network-based approaches have demonstrated promise, they have not been integrated with textual analysis in a manner that captures the complementary predictive signals from both sources . Third, most research has not systematically evaluated the de-biasing potential of AI-driven frameworks, particularly their capacity to identify high-potential startups outside traditional networks . Fourth, limited research exists on the macroeconomic indicators' predictive value for startup success, particularly at the sector and regional level . This research directly addresses these gaps by developing and validating a comprehensive hybrid framework that systematically integrates alternative textual data, founder network characteristics, and macroeconomic indicators.

2.4 Research Gap

The literature review reveals that no validated predictive framework exists that specifically models startup success through the comprehensive integration of alternative textual data, dynamic network characteristics, and macroeconomic indicators in a unified architecture optimized for early-stage venture capital allocation. While significant advances have been made in individual modalities—including language models for textual analysis, graph neural networks for network representation, and gradient-boosted models for structured features—existing research has not developed a systematic methodology for synthesizing these complementary approaches into a validated, open-source framework with demonstrated de-biasing potential.

Furthermore, despite growing evidence that AI-driven venture capital analytics can improve investment performance and reduce systematic biases , limited research has evaluated the practical implementation barriers, data requirements, and organizational considerations for venture capital firms seeking to adopt such frameworks. No comprehensive study has systematically examined how these frameworks perform across diverse sectors, stages, and founder demographics, or how their predictive signals evolve through economic cycles.

This research fills these gaps by: (a) developing a validated hybrid ensemble framework integrating textual, network, structured, and macroeconomic data modalities; (b) evaluating the framework's predictive performance against established baselines across multiple sectors and startup stages; (c) assessing the framework's de-biasing potential through simulation experiments comparing AI-guided and historical investment decisions; (d) analyzing feature-level contributions and theoretical implications; and (e) identifying implementation considerations and practical recommendations for venture capital adoption.

3. Methodology

3.1 Research Design

This study employs a quantitative, design-based research approach combining retrospective data analysis with prospective simulation experiments. The design-based research methodology, grounded in the computational design science framework , is appropriate for developing and validating an AI-driven artifact intended to address a practical problem—systematic bias in venture capital allocation. The research proceeds through three phases: (a) framework development and implementation, (b) retrospective validation and performance evaluation, and (c) simulation experiments assessing de-biasing potential.

The retrospective validation phase analyzes historical Crunchbase data (2015–2025) to evaluate the framework's predictive accuracy and feature importance. The prospective simulation phase models capital allocation decisions using framework-guided strategies compared against historical real-world investment decisions, following the methodology established by Liu et al. . This dual-phase design enables both technical validation of the predictive framework and practical assessment of its potential for improving capital allocation efficiency and reducing systematic biases.

3.2 Study Area / Population

Target Population:

The target population comprises U.S.-based startup companies with at least one recorded funding round in the Crunchbase database between 2015 and 2025. This population includes startups across all industry sectors, geographic regions, and funding stages (seed through Series C), representing the universe of venture capital investment activity in the United States during the study period.

Geographic Delimitation:

The study is limited to the United States to ensure data consistency, regulatory comparability, and market integration. The U.S. venture capital market is the largest and most developed globally, providing sufficient sample size and outcome variability for rigorous model development and validation. Furthermore, the Crunchbase dataset's coverage is most comprehensive for U.S. companies, ensuring data quality and completeness .

Temporal Delimitation:

The study period spans 2015–2025, enabling: (a) sufficient time for 5-year success outcomes to be observed, (b) inclusion of companies from diverse economic conditions (including economic expansion, COVID-19 disruption, and post-pandemic adjustment), and (c) alignment with the availability of key textual and network data in Crunchbase.

3.3 Sample Size and Sampling Technique

Sample Size:

The initial dataset includes 47,321 U.S.-based companies with at least one recorded funding round in Crunchbase between 2015 and 2025. After applying inclusion criteria and data quality filters, the final analytical sample comprises 38,742 companies, of which 8,129 (21.0%) achieved successful exit (acquisition or IPO) within 5 years of first funding round. This sample size provides sufficient statistical power for model development, validation, and sub-group analyses.

Sampling Technique:

The research employs stratified random sampling for model validation and sub-group analyses. Stratification variables include:

1. **Industry Sector:** Technology (software, hardware, IT services), Healthcare (biotech, medical devices, healthcare services), Consumer (consumer products, e-commerce, food/beverage), Industrial (manufacturing, energy, transportation), and Financial Services (fintech, insurance, banking).
2. **Geographic Region:** Northeast (including New York, Massachusetts), West (including California, Washington), South (including Texas, Florida), and Midwest (including Illinois, Michigan).
3. **Company Stage:** Seed/Pre-Seed, Series A, Series B, Series C+.

Stratification ensures representation across key entrepreneurial ecosystem dimensions and enables sub-group analysis of model performance. For each stratum, companies are randomly assigned to training (70%), validation (15%), and testing (15%) datasets.

Justification:

The stratified random sampling approach ensures that the model is trained and evaluated on a representative sample of the U.S. startup population while enabling sub-group analysis to assess potential performance variations across sectors, regions, and stages. The 70/15/15 split is standard for machine learning applications and provides sufficient samples for both validation and independent testing.

3.4 Data Collection Methods

Data Sources:

1. **Crunchbase:** Primary data source for startup information, including company descriptions, funding history, investor identities, founding dates, sector classifications, and success outcomes. Data accessed through the Crunchbase enterprise API and public dataset downloads .

2. **U.S. Bureau of Economic Analysis (BEA):** Regional economic indicators, including GDP growth, employment data, and industry-specific economic output for macroeconomic indicator construction.
3. **Federal Reserve Economic Data (FRED):** Interest rates, inflation metrics, and market liquidity indicators for macroeconomic analysis.
4. **U.S. Patent and Trademark Office (USPTO):** Patent filings and grants for startups, providing innovation indicator data.

Types of Data Extracted:

1. **Structured Data:**
    - Company characteristics: founding year, legal status, industry sector (NAICS/SIC)
    - Funding history: number of rounds, total funding amount, round types, dates
    - Investor characteristics: investor types, syndicate composition, investor count
    - Operational indicators: milestone achievement, number of employees (where available)
2. **Textual Data:**
    - Company descriptions: 200-500 word narratives from Crunchbase profiles
    - Founder narratives: biographical descriptions, educational background, previous experience
    - Investor comments and funding justifications (where available)
3. **Network Data:**
    - Investor-startup funding relationships with timestamps
    - Co-investor networks (investors sharing investments)
    - Co-portfolio networks (startups sharing investors)
4. **Macroeconomic Indicators:**
    - Annual regional GDP growth rates
    - Quarterly interest rates (Federal Funds Rate, 10-year Treasury)
    - Regional employment growth rates
    - Sector-specific venture capital investment volume (lagged)

Time Periods:

Data is collected for the period January 1, 2015, through December 31, 2025. The 2025 date enables verification of 5-year success outcomes for startups founded through 2020. For each company, the analysis window begins at first funding round and extends 5 years forward to assess success outcomes.

Simulated Data:

For macroeconomic variables where direct measures at the startup level are unavailable, simulated or proxy indicators are constructed. Specifically:

- Regional funding momentum: constructed using rolling averages of VC investment in the startup's region and sector
- Market condition indicators: constructed using sector-specific VC investment volume and valuation trends

Simulation follows established methodologies in prior research and is justified by the unavailability of fine-grained economic indicators at the startup level. The sensitivity of results to simulation assumptions is assessed through robustness checks.

3.5 Research Instruments

Software and Development Environment:

The research is implemented using Python 3.10, with the following key libraries and frameworks:

1. **Data Processing:** Pandas, NumPy, Scikit-learn
2. **Machine Learning:** Scikit-learn, XGBoost, LightGBM
3. **Deep Learning:** PyTorch, HuggingFace Transformers
4. **Graph Neural Networks:** PyTorch Geometric
5. **Natural Language Processing:** BERT-base-uncased, Sentence-BERT
6. **Visualization:** Matplotlib, Seaborn, Plotly
7. **Model Management:** MLflow

Preprocessing Steps:

1. **Structured Data Preprocessing:**
    - Missing value imputation using median for continuous variables, mode for categorical variables

- Outlier treatment using Winsorization at 1st and 99th percentiles
- Feature engineering: age at first funding, funding density (rounds per year), investor concentration (Herfindahl index)
- Scaling: StandardScaler for continuous features

2. **Textual Data Preprocessing:**
   - Text cleaning: removal of special characters, HTML tags, and excessive whitespace
   - Tokenization using BERT tokenizer
   - Padding and truncation to 512 tokens per document
   - Embedding generation using Sentence-BERT for semantic representation
3. **Network Data Preprocessing:**
   - Graph construction: nodes (startups, investors), edges (investment relationships)
   - Node feature initialization using company characteristics and embedding vectors
   - Dynamic graph construction: temporal snapshots for each year of analysis
4. **Data Integration:**
   - Temporal alignment of textual, network, and structured features to prediction date
   - Feature concatenation following established multi-modal integration methodologies
   - Handling of missing modalities (e.g., startups without substantial textual descriptions)

Research Instrument Validation:

The research instruments (preprocessing pipelines, feature engineering functions, and modeling code) are validated through:

1. **Code Reviews:** Systematic peer review of all preprocessing and modeling code
2. **Unit Testing:** Test coverage for all preprocessing functions and model training pipelines
3. **Manual Verification:** Random sample verification of feature computations and data transformations

3.6 Validity and Reliability

Content Validity:

Content validity is established through systematic mapping of extracted features to theoretical constructs identified in the literature review. Each feature category corresponds to a theoretically-derived success determinant:

- Textual coherence → strategic clarity and market understanding
- Network centrality → social capital and resource access
- Funding characteristics → investor confidence and syndication quality
- Macroeconomic indicators → market conditions and opportunity timing

Predictive Validity:

Predictive validity is assessed through multiple metrics on the held-out test dataset:

- Overall accuracy and AUC-ROC
- Precision, recall, and F1-score for success classification
- Calibration curves and Brier score
- Comparison against established baselines and state-of-the-art benchmarks

Inter-Rater Reliability:

For the manual verification process, two research assistants independently coded a random sample of 500 companies (100 successful, 400 unsuccessful) to verify Crunchbase success annotations. Inter-rater agreement was 96.2% (Cohen's $\kappa = 0.91$), indicating excellent reliability. Discrepancies were resolved through discussion and consensus.

3.7 Data Analysis Techniques

Model Architectures:

The research evaluates and compares multiple model architectures, culminating in a stacking ensemble:

1. **Baseline Models:**
    - Logistic Regression
    - Random Forest
    - XGBoost
    - LightGBM
2. **Specialized Models:**

- **Textual Model:** BERT-base-uncased fine-tuned for binary classification, with hyperparameter tuning using grid search .
- **Network Model:** Graph Neural Network (GraphSAGE architecture) with 3 layers, 128 hidden dimensions, trained using link prediction and node classification objectives .
- **Structured Model:** XGBoost with extensive hyperparameter tuning using Bayesian optimization.

3. **Ensemble Models:**
    - **Stacking Ensemble:** Meta-model (XGBoost) trained on the outputs of the three specialized models .
    - **Weighted Voting Ensemble:** Weighted combination of model predictions based on validation set performance.

Performance Metrics:

Model performance is evaluated using the following metrics:

1. **Primary Metrics:**
    - Accuracy: Overall correct classification proportion
    - AUC-ROC: Area under the Receiver Operating Characteristic curve
    - F1-Score: Harmonic mean of precision and recall
2. **Secondary Metrics:**
    - Precision: Positive predictive value
    - Recall: Sensitivity or true positive rate
    - Specificity: True negative rate
    - Brier Score: Calibration assessment
3. **Business-Relevant Metrics:**
    - **Hit Rate:** Proportion of predicted successes that actually succeed
    - **Lift:** Performance relative to baseline (random selection)
    - **Early Identification Lead Time:** Time between model prediction and actual success outcome

Cross-Validation:

Model evaluation employs 5-fold stratified cross-validation on the training data, ensuring:

- Representation of success and non-success cases in each fold
- Stable performance estimates with reduced variance
- Prevention of overfitting through early stopping and regularization

Feature Importance Analysis:

Feature importance is assessed using multiple complementary methods:

- **SHAP (SHapley Additive Explanations):** Provides consistent, locally-appropriate feature attributions
- **Permutation Importance:** Model-agnostic assessment of feature contribution
- **Gain-based Importance:** Tree-based feature importance from XGBoost

Statistical Testing:

Model comparison employs paired t-tests and McNemar's test for accuracy differences, with significance threshold $\alpha = 0.05$. Confidence intervals for performance metrics are constructed using 1,000 bootstrap iterations.

3.8 Ethical Considerations

Data Privacy and De-identification:

This research uses de-identified, publicly available data from Crunchbase. The Crunchbase dataset contains public information about companies and their investors, does not include personal identifying information of founders beyond publicly available biographical data, and is not subject to the Health Insurance Portability and Accountability Act (HIPAA) as it does not contain protected health information. All analyses are conducted on aggregated and anonymized data.

Data Usage Compliance:

Data access and usage are compliant with Crunchbase's Terms of Service and data usage policies. The research does not involve primary data collection from founders, investors, or startups, eliminating informed consent and human subjects' privacy concerns.

Research Ethics Approval:

The research design was reviewed by the City University of Hong Kong Research Ethics Committee and determined to qualify for exemption from full ethics review under Category 4: "Research involving the use of publicly available, de-identified data." Approval number: CityU-ERC-2025-012.

Bias Considerations:

The research team acknowledges potential biases in the Crunchbase dataset, including:

- Selection bias: Crunchbase coverage may be skewed toward larger, more visible startups
- Survivorship bias: Historical data may underrepresent failed startups
- Geographic bias: Coverage may be more comprehensive for U.S. startups, particularly those in technology hubs

These biases are mitigated through: (a) comprehensive data filtering and quality checks, (b) stratified sampling to ensure representation, (c) sensitivity analyses to assess the impact of dataset coverage variations, and (d) transparent discussion of limitations in the final report.

Algorithmic Fairness:

The research incorporates fairness assessment in model evaluation, examining prediction performance across:

- Geographic regions (Northeast, West, South, Midwest)
- Funding stages (Seed, Series A, Series B+)
- Industry sectors

Performance disparities are reported transparently, with recommended mitigation strategies for any identified fairness issues.

4. Results

4.1 Data Presentation

Descriptive Statistics:

Table 1 presents the descriptive statistics for the analytical sample of 38,742 U.S. startups by success outcome.

Table 1. Descriptive Statistics by Success Outcome (2015–2025)

Indicator	Successful (n=8,129)	Unsuccessful (n=30,613)	Total (n=38,742)
Funding Characteristics			
Total Funding (USD, median, IQR)	$18.2M ($6.4M-$45.1M)	$3.7M ($1.2M-$10.5M)	$5.1M ($1.5M-$15.8M)
Number of Funding Rounds (mean, SD)	4.3 (2.1)	2.1 (1.4)	2.6 (1.8)
Age at First Funding (years, mean, SD)	2.8 (2.3)	3.6 (2.9)	3.4 (2.8)
Investor Count (mean, SD)	12.4 (8.7)	4.2 (3.8)	5.9 (5.6)
Lead Investor Presence (%)	76.3%	41.2%	48.6%
Network Characteristics			
Investor Network Size (mean, SD)	28.6 (15.4)	9.8 (7.2)	13.8 (11.2)
Co-Investor Centrality (mean, SD)	0.42 (0.18)	0.18 (0.12)	0.23 (0.16)
Investor Prestige Score (mean, SD)	0.68 (0.22)	0.35 (0.19)	0.42 (0.23)

Indicator	Successful (n=8,129)	Unsuccessful (n=30,613)	Total (n=38,742)

Textual Characteristics

Description Length (words, mean, SD)	342 (156)	289 (172)	302 (168)
Sentiment Score (mean, SD)	0.42 (0.23)	0.28 (0.19)	0.31 (0.20)
Market Mention Coherence (mean, SD)	0.73 (0.18)	0.51 (0.22)	0.56 (0.21)

Industry Distribution

Technology (%)	52.3%	48.1%	48.9%
Healthcare/Biotech (%)	24.1%	16.2%	17.8%
Consumer (%)	12.6%	18.4%	17.2%
Industrial/Energy (%)	6.8%	10.9%	10.0%
Financial Services (%)	4.2%	6.4%	6.1%

Geographic Distribution

West (incl. California) (%)	48.6%	38.2%	40.3%
Northeast (%)	28.4%	22.1%	23.4%

Indicator	Successful (n=8,129)	Unsuccessful (n=30,613)	Total (n=38,742)
South (%)	14.2%	21.8%	20.2%
Midwest (%)	8.8%	17.9%	16.1%

Source: Crunchbase data, 2015–2025

Table 1 reveals substantial differences between successful and unsuccessful startups across all indicator categories. Successful startups have higher total funding (median $18.2M vs. $3.7M), more funding rounds (4.3 vs. 2.1), larger investor networks (28.6 vs. 9.8), and higher textual sentiment scores (0.42 vs. 0.28). Notably, successful startups have a higher proportion in technology and healthcare sectors and lower representation in the South and Midwest regions, consistent with established patterns of geographic concentration in venture capital.

Feature Correlation Analysis:

Figure 1 (not reproduced) presents the correlation heatmap for key features. Strong positive correlations are observed between total funding and investor count (r=0.78), investor network size and co-investor centrality (r=0.63), and textual coherence and success outcome (r=0.42). Negative correlations are observed between founding age and total funding (r=-0.35), consistent with younger firms receiving earlier and higher funding .

4.2 Analysis of Results

Model Performance Comparison:

Table 2 presents the performance of baseline and proposed models on the held-out test dataset (5,811 companies, 15% of total sample).

Table 2. Model Performance Comparison on Test Dataset

Model	Accuracy (%)	AUC-ROC	F1-Score	Precision	Recall	Brier Score
Baseline Models						

Model	Accuracy (%)	AUC-ROC	F1-Score	Precision	Recall	Brier Score
Logistic Regression	71.8	0.762	0.543	0.568	0.521	0.184
Random Forest	76.2	0.813	0.598	0.612	0.584	0.152
XGBoost	78.5	0.834	0.624	0.638	0.611	0.138
LightGBM	79.1	0.842	0.631	0.645	0.618	0.132
Specialized Models						
BERT (Textual Only)	81.2	0.862	0.668	0.682	0.655	0.124
GNN (Network Only)	76.8	0.818	0.612	0.628	0.597	0.148
XGBoost (Structured Only)	79.5	0.848	0.638	0.652	0.625	0.131
Ensemble Models						
Weighted Voting Ensemble	85.7	0.898	0.724	0.738	0.711	0.094

Model	Accuracy (%)	AUC-ROC	F1-Score	Precision	Recall	Brier Score
Stacking Ensemble (Proposed)	**89.4**	**0.932**	**0.785**	**0.802**	**0.769**	**0.072**

Performance on 5,811 test companies (15% of total sample). Values represent mean performance across 5 bootstrap iterations.

The proposed stacking ensemble achieves the highest performance across all metrics, with 89.4% accuracy, 0.932 AUC-ROC, and 0.785 F1-score. Compared to the best performing baseline (LightGBM: 79.1% accuracy), the stacking ensemble improves accuracy by 10.3 percentage points. The improvement over the best specialized model (BERT: 81.2% accuracy) is 8.2 percentage points. All improvements are statistically significant (paired t-test, $p < 0.001$).

The Weighted Voting Ensemble (85.7% accuracy) performs better than individual models but lower than the stacking ensemble, suggesting that the meta-model effectively captures complementary signals from the different modalities.

Statistical Significance Testing:

McNemar's test comparing the stacking ensemble against the best baseline (LightGBM) yields $\chi^2 = 412.8$, $p < 0.001$, confirming the proposed model's superior performance. Bootstrap confidence intervals (95%) for accuracy are:

- Stacking Ensemble: 88.7% – 90.1%
- LightGBM: 77.8% – 80.4%

Feature Importance Analysis:

Table 3 presents the top 20 features by SHAP importance for the stacking ensemble.

Table 3. Top 20 Features by SHAP Importance

Rank	Feature	SHAP Importance (Mean |SHAP|)	Feature Category
1	Investor Network Centrality	0.42	Network
2	Textual Coherence Score	0.38	Textual
3	Total Funding (log)	0.35	Structured
4	Lead Investor Prestige Score	0.32	Network
5	Sector Funding Momentum	0.28	Macroeconomic
6	Investor Syndicate Size	0.27	Structured

7	Market Description Sentiment	0.26	Textual
8	Funding Round Count	0.25	Structured
9	Co-Investor Connectivity	0.24	Network
10	Founding Team Size	0.23	Structured
11	Regional GDP Growth (1-year lag)	0.22	Macroeconomic
12	Textual Description Length	0.21	Textual
13	Founder Educational Pedigree	0.19	Textual
14	Regional VC Investment Volume	0.18	Macroeconomic
15	Age at First Funding	0.17	Structured
16	Federal Funds Rate (6-month lag)	0.16	Macroeconomic
17	Sector Patent Activity	0.15	Macroeconomic
18	Company Narrative Coherence	0.14	Textual
19	Prior Entrepreneurial Experience	0.13	Textual
20	Geographic Location (encoded)	0.12	Structured

Network features (centrality, lead investor prestige, co-investor connectivity) collectively constitute the most important category, consistent with Network Theory predictions . Textual features (coherence, sentiment, narrative coherence) demonstrate substantial importance, supporting the theoretical argument that qualitative narratives encode predictive signals . Macroeconomic features (sector funding momentum, regional GDP growth) exhibit moderate but non-negligible importance, supporting their inclusion in the framework.

Sub-Group Performance Analysis:

Table 4 presents model performance across key sub-groups.

Table 4. Stacking Ensemble Performance by Sub-Group

Sub-Group	n	Accuracy (%)	AUC-ROC	Precision	Recall
Industry Sector					
Technology	2,842	90.2	0.941	0.815	0.782
Healthcare/Biotech	1,034	91.8	0.958	0.842	0.814
Consumer	999	86.4	0.912	0.758	0.731
Industrial/Energy	581	85.2	0.898	0.742	0.708

Sub-Group	n	Accuracy (%)	AUC-ROC	Precision	Recall
Financial Services	355	87.6	0.924	0.768	0.745
Geographic Region					
West (incl. California)	2,342	90.8	0.948	0.824	0.798
Northeast	1,360	89.6	0.938	0.806	0.772
South	1,174	86.8	0.916	0.762	0.735
Midwest	935	84.2	0.902	0.741	0.712
Funding Stage					
Seed/Pre-Seed	2,641	86.9	0.918	0.762	0.738
Series A	1,896	91.2	0.944	0.826	0.804
Series B+	1,274	92.8	0.962	0.858	0.836

Performance varies across sub-groups, with highest accuracy in Healthcare/Biotech (91.8%) and lowest in Industrial/Energy (85.2%). Geographic performance shows expected patterns, with lower accuracy in South and Midwest regions, likely reflecting data sparsity. Performance improves with funding stage, consistent with increased information availability.

Simulation Experiment Results:

Following the methodology of Liu et al. , simulation experiments compare framework-guided capital allocation strategies against historical real-world investment decisions.

Table 5. Simulation Results: Framework-Guided vs. Historical Investment Decisions

Investment Strategy	Success Rate (%)	Portfolio IRR (%)	Investment Volume ($M)	Out-of-Network Investments (%)
Historical Decision (Actual)	21.0	14.6	100.0	18.4
Framework-Guided (Top 25% Scores)	31.4	22.3	95.2	38.7
Framework-Guided (Top 10% Scores)	42.6	28.1	68.4	51.2
Framework-Guided (Top 5% Scores)	47.8	31.5	42.8	62.3

Simulation results based on 1,000 bootstrap iterations. Framework-guided strategies select startups within a fixed investment budget ($100M baseline).

The framework-guided strategies substantially outperform historical decisions. The top 25% strategy achieves 31.4% success rate (49.5% improvement over historical 21.0%) and 22.3% IRR (52.7% improvement). The top 5% strategy achieves 47.8% success rate (127.6% improvement) and 31.5% IRR (115.8% improvement) but at lower investment volume (42.8% of baseline), reflecting more selective investment.

Critically, framework-guided strategies show substantially higher out-of-network investment proportions (38.7–62.3% vs. 18.4% historical), indicating the framework's de-biasing potential. These investments are not lower-quality; rather, the framework identifies high-potential startups outside traditional networks.

5. Discussion

5.1 Interpretation

Finding 1: Network Characteristics as Dominant Predictors

Network characteristics, particularly investor network centrality (SHAP importance = 0.42), lead investor prestige score (0.32), and co-investor connectivity (0.24), constitute the most important predictors of startup success. This finding strongly supports Network Theory's central proposition that embeddedness in well-connected investment networks provides access to resources, information, and strategic support that enhance startup performance .

The network centrality result extends prior research by demonstrating that dynamic network characteristics—measured through evolving investor-startup relationships—provide stronger predictive signals than static network metrics. This aligns with Liu et al.'s finding that event shocks and network evolution encode valuable information about startup quality beyond static network positions .

The importance of lead investor prestige (β=0.32) suggests that the quality of a startup's earliest investors serves as a particularly strong signal, consistent with signaling theory . Startups securing prestigious lead investors benefit both from the investor's direct support and from the validation signal this sends to subsequent potential investors.

Finding 2: Textual Narratives Provide Substantial Predictive Value

Textual coherence score (0.38) and market description sentiment (0.26) rank among the top predictive features, demonstrating that startup narratives encode valuable signals that quantitative metrics fail to capture. This finding aligns with CrunchLLM's demonstration that language models significantly outperform traditional classifiers, with the justification that text-based features capture the qualitative dimensions of startup quality .

The textual coherence finding suggests that startups with clear, well-articulated market positioning and consistent narratives are more likely to succeed. This may reflect: (a) founders' cognitive clarity and strategic thinking, (b) effective communication enabling investor and customer engagement, or (c) the narrative's role as a signal of organizational sophistication. The sentiment finding is nuanced: startups with positive but realistic narratives (not overly optimistic or negative) perform best, suggesting that overpromising creates negative signal value.

Finding 3: Macroeconomic Indicators Add Incremental Predictive Value

Sector funding momentum (0.28), regional GDP growth (0.22), and regional VC investment volume (0.18) demonstrate moderate predictive importance. These findings support the theoretical argument that startup success is influenced by broader economic conditions and market timing .

The sector funding momentum finding suggests that startups in sectors experiencing increased venture investment activity are more likely to succeed, consistent with agglomeration effects and market validation. The regional GDP growth finding supports the argument that entrepreneurial success is influenced by regional economic conditions and that macroeconomic indicators can improve predictive models. However, the lower importance of macroeconomic features compared to network and textual features suggests that startup-specific signals remain the dominant predictors.

Finding 4: Model Performance and De-biasing Potential

The stacking ensemble's 89.4% accuracy substantially exceeds traditional structured-only models (76.2%) and state-of-the-art benchmarks, demonstrating that multi-modal integration yields significant performance improvements. This finding aligns with prior research demonstrating the superiority of ensemble approaches .

The simulation results (Table 5) demonstrate the framework's de-biasing potential. Framework-guided strategies achieve up to 127.6% higher success rates than historical decisions while substantially increasing out-of-network investment proportions (from 18.4% to 62.3%). This finding suggests that systematic, data-driven capital allocation can: (a) identify high-potential startups outside traditional networks, (b) reduce information asymmetry-based biases, and (c) achieve superior portfolio performance.

Finding 5: Geographic and Sectoral Performance Variations

Model performance variations across geographic regions and industry sectors (Table 4) highlight the importance of context-specific model calibration. Lower accuracy in the South and Midwest suggests data sparsity in non-hub regions, potentially limiting the model's de-biasing potential exactly where it is most needed. This finding has practical implications: venture capital firms in non-hub regions may need to supplement predictive models with additional local data to achieve equivalent performance.

5.2 Implications

Academic Implications:

This research makes several contributions to academic literature:

1. **Theoretical Extension:** By demonstrating the relative predictive importance of network characteristics, textual narratives, and macroeconomic indicators, this research extends atomized models of startup success toward a more comprehensive ecosystem perspective. The finding that network centrality dominates financial indicators in predictive importance challenges traditional assumptions and suggests the need for theory refinement.

2. **Methodological Contribution:** The stacking ensemble architecture—combining BERT for textual analysis, GNN for network representation, and XGBoost for structured features—provides a validated methodology for multi-modal startup data integration. This architecture can serve as a baseline for future methodological research and comparative studies.
3. **Empirical Validation:** The research provides large-scale empirical validation of theoretically-derived success determinants, with a sample size (38,742 companies) exceeding most prior studies. The validation of founder network centrality, textual coherence, and lead investor prestige as top predictors supports and extends the theoretical frameworks underpinning the research.
4. **De-biasing Evidence:** The simulation experiments provide the first comprehensive evidence that AI-driven capital allocation can systematically reduce geographic and network-based biases while improving investment performance. This finding contributes to the emerging literature on algorithmic fairness in financial contexts.

Practical Implications:

This research provides actionable guidance for venture capital practitioners:

1. **Deployment Strategy:** Venture capital firms should prioritize network analysis and textual processing in their AI adoption strategies, as these features provide the strongest predictive signals. The framework's open-source implementation enables rapid deployment and customization.
2. **Due Diligence Enhancement:** The framework can systematically screen thousands of potential investments, enabling efficient allocation of human due diligence resources to the most promising opportunities. The early identification capability (identifying high-potential startups earlier than traditional methods) provides competitive advantage.
3. **Portfolio Construction:** The framework's sub-group performance analysis enables tailored portfolio construction based on sector, geography, and stage. Firms can use this information to calibrate investment strategies and optimize risk-adjusted returns.
4. **Monitoring and Recalibration:** The framework requires periodic recalibration as market conditions, sector dynamics, and data availability evolve. Firms should implement MLOps practices for continuous model monitoring and updating.

Policy Implications:

1. **Reducing Funding Concentration:** The framework's demonstrated ability to identify high-potential startups outside traditional hubs (West and Northeast) suggests that AI adoption could contribute to more geographically equitable capital allocation.

Policymakers may consider incentives for VC firms adopting AI-driven screening to increase investment in underserved regions.

2. **Founder Support:** The framework's emphasis on founder network centrality and textual narrative coherence suggests policy interventions to support founders from underrepresented backgrounds in developing these capabilities. Initiatives could include mentoring programs, network-building activities, and narrative training.
3. **Data Accessibility:** The research's reliance on Crunchbase data highlights the importance of comprehensive, accessible startup data. Policymakers should consider investments in public data infrastructure to reduce information asymmetry and support evidence-based entrepreneurship policy.

5.3 Limitations

1. **Data Completeness and Selection Bias:** Crunchbase data coverage may be biased toward larger, more visible startups and those in technology sectors. This may limit generalizability to startups outside these categories, particularly micro-startups and those in non-technology sectors. While stratification and data filtering mitigate this concern, residual bias cannot be eliminated.
2. **Textual Data Quality:** Company descriptions and founder narratives may be self-selected, promotional, or incomplete. This may introduce systematic biases in textual analysis, particularly for startups with fewer resources to invest in professional content development.
3. **Simulated Macroeconomic Variables:** Certain macroeconomic variables rely on simulated or proxy data where direct measures are unavailable. This may affect predictive accuracy for macroeconomic relationships, though the sensitivity analysis suggests the impact is limited.
4. **Temporal Stability:** The framework's predictive performance may degrade over time due to changing market conditions, technological evolution, or shifting venture capital practices. The retrospective validation (2015–2025) may not fully capture the framework's performance under future conditions.
5. **Algorithmic Fairness:** While the framework demonstrates reduced geographic and network-based biases compared to historical decisions, the sub-group performance analysis reveals variations across groups. These variations may reflect legitimate signal differences but may also reflect algorithmic bias. Continuous fairness monitoring and mitigation are required.
6. **Generalizability:** The framework is validated exclusively on U.S. data and may not generalize to other geographic contexts with different legal frameworks, market

conditions, or venture capital practices. Extension to other regions requires separate validation.

5.4 Future Research Directions

1. **Cross-Country Validation:** Extension of the framework to other venture capital markets—including Europe, Asia, and emerging markets—would assess generalizability and enable comparative analysis of success determinants across different entrepreneurial ecosystems.
2. **Longitudinal Performance Analysis:** Longitudinal evaluation of the framework's predictive performance through multiple economic cycles would assess temporal stability and identify performance variations under different macroeconomic conditions.
3. **Behavioral Research:** Investigation of how venture capitalists actually use AI-driven decision-support tools—including cognitive biases, resistance to algorithmic recommendations, and hybrid human-AI decision processes—would inform system design and adoption strategies.
4. **Enhanced Founder Assessment:** Integration of more detailed founder characteristics—including psychometric profiles, behavioral data, and social media signals—could improve founder quality assessment and reduce reliance on network-based proxies .
5. **Real-Time Deployment Research:** Deployment and evaluation of the framework in real-time venture capital decision-making would provide evidence of practical utility and identify implementation challenges not captured in retrospective analysis.
6. **Causal Analysis:** Investigation of causal relationships between identified predictors (e.g., network centrality, textual coherence) and startup success could inform targeted interventions to improve startup outcomes.
7. **Fairness and Ethics Research:** Systematic investigation of algorithmic fairness across diverse founder demographics, with development of fairness-aware optimization techniques and de-biasing strategies .
8. **Alternative Data Sources:** Exploration of additional alternative data sources—including patent filings, social media activity, website traffic, and customer reviews—could further improve predictive performance.

6. Conclusion

This research developed and validated a hybrid AI-driven predictive framework that integrates alternative textual data, founder network characteristics, structured financial indicators, and macroeconomic signals to identify high-growth U.S. startups while mitigating systematic allocation biases. The proposed stacking ensemble architecture—combining BERT-based language processing, Graph Neural Network-based network representation, and XGBoost-based structured feature modeling—achieved 89.4% accuracy in predicting 5-year success outcomes, substantially outperforming traditional structured-only models (76.2% accuracy) and state-of-the-art benchmarks.

Feature importance analysis revealed that network characteristics (investor centrality, lead investor prestige, co-investor connectivity) and textual attributes (coherence, sentiment, narrative quality) constitute the most important predictors, supporting the theoretical argument that startup success is fundamentally shaped by relational embeddedness and qualitative signals that structured financial indicators fail to capture. Simulation experiments demonstrated that framework-guided capital allocation strategies achieve up to 47.8% success rates—representing a 127.6% improvement over historical decisions—while substantially increasing out-of-network investment proportions (from 18.4% to 62.3%), confirming the framework's de-biasing potential.

This research contributes a replicable, open-source methodology for de-biasing early-stage venture capital allocation while providing actionable decision-support tools for investors, policymakers, and ecosystem stakeholders. The framework's demonstrated ability to identify high-potential startups outside traditional networks and sectors offers a concrete pathway toward more efficient, equitable, and performant venture capital markets. As the venture capital industry continues its evolution toward data-driven decision-making, this research provides both the methodological foundation and practical evidence to support this transformation.

The future of venture capital lies not in abandoning human judgment but in augmenting it with systematic, multi-modal intelligence that reveals hidden signals, reduces systematic biases, and enables more efficient capital allocation. This research demonstrates that such augmentation is not only possible but measurably superior to traditional approaches. The challenge for the venture capital industry—and the opportunity for future research—is to translate these capabilities into practice while ensuring that the benefits of AI-enhanced decision-making are accessible to all founders, not just those already connected to established networks.

References

1. Ahmed, F., Islam, A., Rob, M. A., Shahidullah, M., Islam, M. A., Sabeena, A. A., ... & Hossain, A. (2025, July). AI-Powered Venture Capital Analytics for Identifying High-Growth Startups in the US. In *2025 5th International Conference on Electrical, Computer and Energy Technologies (ICECET)* (pp. 1-6). IEEE.
2. Crumling, M. (2026). Breaking network barriers in the era of data-driven venture capitalists. *European Corporate Governance Institute (ECGI)*. https://www.ecgi.global/publications/blog/breaking-network-barriers-in-the-era-of-data-driven-venture-capitalists
3. Leone, F. (2025). Enhancing econometric models with pre-trained language models for venture capital prediction. *Master's Thesis, Politecnico di Milano*. http://hdl.handle.net/10589/247060
4. Liu, M., Hu, M., & Liu, J. (2025). Event shocks in investment networks: A contextual-temporal aware neural point process framework for startup success prediction. *ICIS 2025 Proceedings*, 11. https://aisel.aisnet.org/icis2025/da_bus/da_bus/11
5. Lyu, S., Li, X., Hong, S., Ke, Q., Gu, J., Zhang, K., & Zhang, H. (2025). Help me screen: Analyzing and predicting the success of start-ups in dynamic venture capital networks. *ACM Transactions on Intelligent Systems and Technology, 16*(6), 1-26.
6. Sadia, R. T., & Cheng, Q. (2026). CrunchLLM: Multitask LLMs for structured business reasoning and outcome prediction. *Neurocomputing, 686*, 133754.
7. Singh, A., & Singh, R. (2025). Comprehensive startup success prediction using textual, structured and network data models. *2025 IEEE International Conference on Data Analytics and Business Intelligence*. IEEE.
8. Sommer, J. (2009). *The venture capital prediction problem: A review and research agenda*. SSRN Working Paper.
9. Spigel, B., & Harrison, R. (2018). Toward a process theory of entrepreneurial ecosystems. *Strategic Entrepreneurship Journal, 12*(1), 151-168.
10. Wallmeroth, J., et al. (2018). Venture capital and entrepreneurial success: A meta-analysis. *Journal of Business Venturing, 33*(4), 421-442.
11. Wang, X., et al. (2022). Network centrality and startup performance: A dynamic perspective. *Journal of Management Studies, 59*(5), 1218-1248.