Using Deep Learning to Find the Next Unicorn: A Practical Synthesis on Optimization Target, Feature Selection, Data Split and Evaluation Strategy

Lele Cao[1], **Vilhelm von Ehrenheim**[1], **Sebastian Krakowski**[2], **Xiaoxue Li**[3], **Alexandra Lutz**[1]
[1] Motherbrain, EQT Group, Stockholm, Sweden
[2] House of Innovation, Stockholm School of Economics, Stockholm, Sweden
[3] Department of Political Science, Stockholm University, Stockholm, Sweden
{lele.cao, vilhelm.vonehrenheim, alexandra.lutz}@eqtpartners.com
sebastian.krakowski@hhs.se, xiaoxue.li@statsvet.su.se

Abstract

Startups represent newly established business models associated with disruptive innovation and high scalability, hence strongly propel the economic and social development. Meanwhile, startups are heavily constrained by many factors such as limited financial funding and human resources. Therefore, the chance for a startup to succeed is rare like "finding a unicorn in the wild". Venture Capital strives to identify and invest in unicorn startups as early as possible, hoping to gain a high return. This work is traditionally manual and empirical, making it inherently biased and hard to scale. Recently, the rapid growth of data volume and variety is quickly ushering in deep learning (DL) as a potentially superior approach in this domain. In this work, we carry out a literature review and synthesis on DL-based approaches, emphasizing four key aspects: optimization target, feature selection, data split, and evaluation strategy. For each aspect, we summarize our in-depth understanding and practical learning.

1 Introduction

Startup is a dynamic, flexible, high risk, and newly created company that typically represents a reproducible and scalable business model. It provides innovative products and/or services, and has limited financial funds and human resources (Santisteban et al., 2021; Skawińska and Zalewski, 2020; Blank, 2013). Since startups stimulate growth, generate jobs and tax revenues, and promote many other socioeconomically beneficial factors (Acs and Szerb, 2007), they are commonly regarded as powerful engines for economic and social development. As the startups continue to develop, they often increasingly rely on external funds (as opposed to internal funds from founders and co-founders), from either domestic or foreign capital markets, to unlock a high rate of growth (Marmer et al., 2011). Up till this date, the dominating external fund source has been Venture Capital (VC).

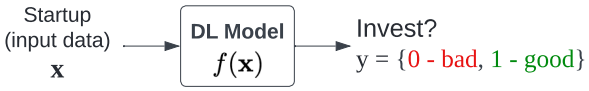


Figure 1: High-level overview of DL (deep learning) based startup sourcing: the model is trained to approximate a function $f(\cdot)$ so that the input $\mathbf{x}$ describing a startup is mapped to an output y indicating the recommended investment propensity that can be either discrete (good vs. bad) or continuous (success probability).

As an industry, VC seeks opportunities to invest in startups with great potential (in the sense of financial returns) to grow and successfully exit. The risk-return trade-off tells us that the potential return rises with a corresponding increase in risk[1]. As a consequence, VC firms strive to mitigate this risk by improving their 1) *deal sourcing* and 2) *value-add* process (Teten et al., 2013). In this survey, we will focus on the published work around the former approach, i.e., finding the *startup unicorn*[2] as accurately as possible. However, this task is a complex one with great uncertainty because of many factors such as vague/immature business ideas, forcing VC firms to make investment decisions based on insufficient information. Therefore a VC's deal sourcing process traditionally turns out to be manual and empirical, leaving estimations of the ROI (return on investment) heavily dependent on the human investors' decisions, which are inherently biased and hard to scale (Cumming and Dai, 2010).

With the rapid growth of data size and diversity (origin and modality), DL (deep learning) methods caught the eyes of increasing number of researchers hunting for unicorns. DL, by definition, represents a subset of ML (machine learning) methods, and is implemented (entirely or partly) with

[1]Statistics revealing the high risk of startups: on average, only about 60% of the startups survive for over 3 years since founded (Hyytinen et al., 2015); top 2% of VCs receive 95% of the returns in the entire industry (Bai and Zhao, 2021); VC has only 10% rate of achieving an ROI (return on investment) of 100% or more (Shane, 2012; Ünal and Ceasu, 2019).

[2]Unicorn startups are private, VC-backed firms with a valuation of at least $500 million (Chernenko et al., 2021).

Proceedings of the Joint Workshop of the 5th Financial Technology and Natural Language Processing (FinNLP) and 2nd Multimodal AI For Financial Forecasting (Muffin), Macao, August 20, 2023.

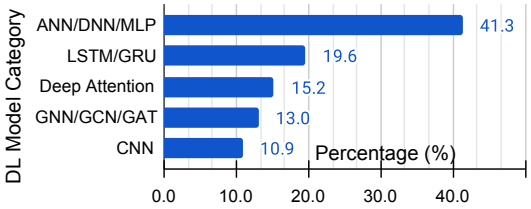


Figure 2: The the adoption percentage of DL models.

ANNs (artificial neural networks) that utilize at least two hidden layers of neurons. As shown in Figure 1, DL-based approaches require practitioners to define the input data $\mathbf{x}$ and label y (indicating good or bad investment according to some criteria) before training a model $f(\cdot)$ that maps $\mathbf{x}$ to y, i.e., $y{=}f(\mathbf{x})$. As a well-known international investment firm practicing data-driven approaches to find startup unicorns, we strive to 1) obtain a thorough and in-depth understanding of the methodologies for startup evaluation using DL, and 2) distil important and actionable learning for practitioners.

Therefore, we carry out a literature survey[3] on using DL to evaluate startups. According to Figure 2, over 40% of the surveyed papers adopt an ANN/DNN/MLP[4] due to its wide applicability to many data types. LSTM/GRU[5] almost dominates the cases when time-series are used. Deep attention (Vaswani et al., 2017) and graph based models (GNN/GCN/GAT)[6] have a rising trend of adoption due to increasing introduction of text and graph input. Lastly, images and videos are relatively least used (Figure 4), leading to only around 10% adoption rate for CNN (convolutional NN). We discover that **the innovation mostly lies in how an existing DL model is applied, rather than in the model itself**. Particularly, we present our literature synthesis and practical learnings from four key aspects: **optimization target**, **feature selection**, **data split**, and **evaluation strategy**. To the best of our knowledge so far, our work is the first of this kind.

2 Optimization Target

Identifying potential unicorns relies on accurate prediction of startup success. So far there is no universally agreed definition of "true success"; most of the existing definitions commonly focus on "growth" which can be measured from different perspectives like revenue, employees, and valuation, to name a few. We summarize the definitions adopted by the reviewed literature, showing each criterion's popularity among researchers. All *success criteria* are quantities in relation to a predefined duration since the time point of evaluation.

1. **Fulfill the preset fundraising goal** (Lee et al., 2018; Yu et al., 2018; Cheng et al., 2019; Yeh and Chen, 2020; Shi et al., 2021; Kaminski and Hopp, 2020; Wu et al., 2022; Tang et al., 2022): the goal (the expected amount of money) of the fund-raise campaign or plan is reached or surpassed, which is common among crowdfunding projects. The readers should be cautious not to confuse with the fund-raise goal of investors.
2. **Future funding** (Chen et al., 2021; Ross et al., 2021; Stahl, 2021; Yin et al., 2021; Garkavenko et al., 2022): any future funding raised above a low-bar amount.
3. **Acquired** (Ang et al., 2022; Ferrati et al., 2021; Kim et al., 2020; Lyu et al., 2021): one company purchases and takes over the operations and assets of the startup.
4. **IPO** (initial public offering) (Ang et al., 2022; Ferrati et al., 2021; Yin et al., 2021): it offers shares to the public in a new stock issuance for the first time; IPO allows the company to raise equity capital from public investors.
5. **Series A** (Zhang et al., 2021; Dellermann et al., 2021): the startup receives the first VC funding round after the seed and angel rounds.
6. N**-year survival** (Ghassemi et al., 2020; Ross et al., 2021): the firm still operates after N years.
7. **Experts view** (Bai and Zhao, 2021; Kinne and Lenz, 2021): the quantified review from human experts.
8. **Upround** (Ang et al., 2022): the valuation after a future funding round becomes higher than the current valuation.
9. **VC-backed** (Garkavenko et al., 2021): the startup is funded by one or more VC firms.
10. **Total raised funding** (Kim and Park, 2017): the accumulated amount of funding received (the higher the better), which is often used as a regression target.

[3]There are 29 English papers/theses sourced (with no restriction of year, type or geo-location) from 1) investment professionals and researchers, 2) keywords searching in Google, Google Scholar, IEEE, ACM, Scopus, Wiley, Springer and Web of Science, and 3) cross reference among papers/theses.

[4]In this paper, ANN, DNN (deep NN) and MLP (multilayer perceptron) refer to a NN with at least two hidden layers.

[5]LSTM: long short term mem.; GRU:gated recurrent unit.

[6]GNN (graph NN), GCN (graph convolution net) and GAT (graph attention net) are three graph based DL models.

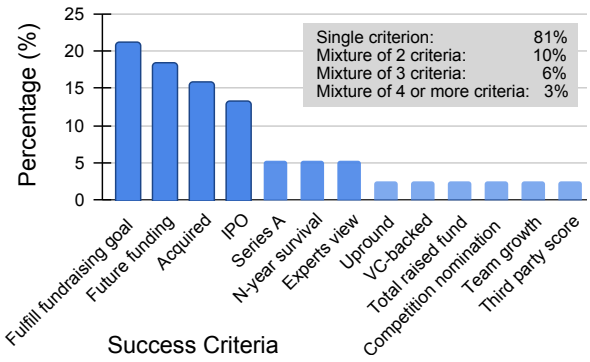


Figure 3: Distribution of the adopted startup success criteria (i.e., optimization objective); the upper-right panel shows the percentage of combining different number of criteria together.

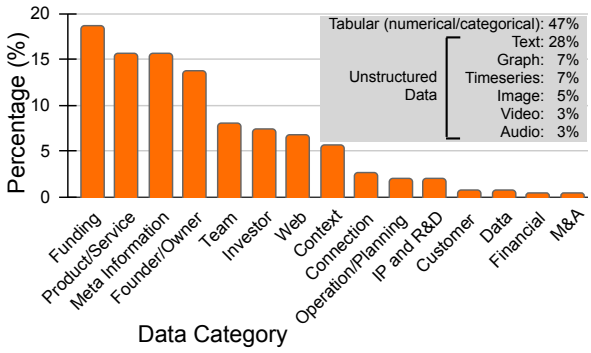


Figure 4: The distribution of data category sorted by the their occurrences; the upper-right panel shows a snapshot (to the date when this paper is written) of the utilized data modalities: numerical, categorical, text, graph, time-series, image, video and audio.

11. **Competition nomination** (Ghassemi et al., 2020): the startup's business idea wins (or nominated by the committee of) a entrepreneurial competition.
12. **Team growth** (Horn, 2021): whether the team size has experienced a fast growth or not, such as "$\geq x\%$ *increase from at least 10 initial employees*".
13. **3rd-party score** (Allu and Padmanabhuni, 2022): some data sources provide certain firm evaluation scores[7].

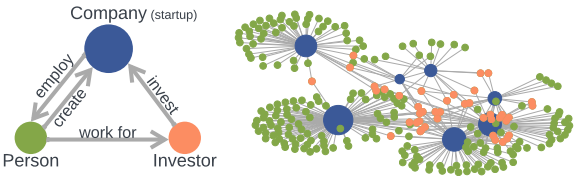
While the first 12 criteria are intuitively sound, we question the effectiveness of the last criterion of taking the 3rd-party (algorithmic) scores as ground truth to train the DL model, because it is guaranteed to obtain a model inferior to the 3rd-party method. Additionally, there is no financial based success criteria[8] adopted in the DL-based work, which is a consequence of missing rich operating data (Gompers et al., 2020) before exiting the startup phase and entering the *growth phase* (Skawińska and Zalewski, 2020). Although the definition of a successful startup has many versions, for investors, it is relatively straightforward: a profitable exit, often in the form of acquisition or IPO, which incur high ROI (Ang et al., 2022). Practically, short-term events like funding rounds have a higher adoption rate than longer-term acquisition/IPO; the reason is twofold: 1) acquisition/IPO is extremely scarce as very few startups achieve these milestones; and 2) it occurs very late in startup's trajectory, hence potentially weakening the correlation between early data and late success (Stahl, 2021). In most cases, different success criteria do not conflict with each other, implying the possibility to combine multiple criteria; but this kind of *criteria mixture* is still under-investigated as illustrated in the upper-right panel of Figure 3. Generally speaking, one can combine multiple criteria with logical operators (i.e., `OR` and `AND`) (Yin et al., 2021; Ang et al., 2022), or use each criterion separately in a multi-task training setup (Shi et al., 2021).

[7]For example, Crunchbase (`www.crunchbase.com`) provides a so called "trend score" score.

[8]Only a few ML-based (instead of DL-based) work (Lussier and Pfeifer, 2001; Lussier and Halabi, 2010) have investigated using financial based success criteria.

3 Feature Selection

DL models need data input to make predictions. Before we start gathering input data for model, we might be able to benefit from understanding what input(s) humans use to make decisions. When investment professionals (i.e., humans) try to forecast the success of early stage startups, they make use of two cognitive modes: *intuitive* and *analytical*. The *intuitive mode* is characterized by processing "soft" signals (e.g., innovativeness and personality of entrepreneur) that are mostly *qualitative*; and humans are still the "golden standard" for this mode (Baer and McKool, 2014). The *analytical mode*, on the other hand, deals with "hard" facts (e.g., industry and team size) that are often *quantitative* (Dellermann et al., 2021). The majority of the work we reviewed incorporate both modes into the model input, but they have to quantify the "soft" information via either approximation or questionnaire. Data is often fed into DL models in the form of *feature*s. **Feature** (a.k.a. "factor" in the scope of financial research) is an individual measurable property or characteristic of a phenomenon, which is sometimes aggregated from raw data. When we try to map out the large number of features used the literature, we found that features tend to cluster into



(a) Entity connections. (b) An example graph.

Figure 5: Illustration of the connection feature category, from which a graph can be built: (a) the graph comprises many nodes (denoting company, person or investor) and edges (representing investing/employment/founding relations between nodes); (b) an example company-person-investor graph using (a) as a legend.

different categories, describing different aspects of the startup in scope. We identified 15 feature categories detailed below and visualize their adoption percentage in Figure 4. Refer to Table 1 for the concrete features adopted in each category.

- **Funding**: historical fund received by the startup is direct recognition from other investors, thus it is the most popular category in the literature.
- **Product/Service**: the core value that early startups have to offer is reflected in the product and/or service they aim to create, which makes this category widely adopted.
- **Meta Information**: the general attributes of startups, which seldomly change since creation.
- **Founder/Owner**: the attributes of founding teams and the individuals that comprise them contribute to both the short-term success and longer-term survival (Ghassemi et al., 2020) of the startup; this category is available from many data sources and entrepreneurial competitions.
- **Team**: complementary to the founder/owner feature, the team features capture the statistics of the employees.
- **Investor**: the statistics of investors that have funded the startup can be informative about its early attractiveness.
- **Web**: any feature extracted from web pages.
- **Context**: besides *intrinsic*[9] features, more and more researchers have realized the importance of *extrinsic* factors that may be (but not limited to) competition, environmental, cultural, economical and tax-based.
- **Connection** features, as illustrated in Figure 5, are usually extracted from a graph that encodes connections between different entities: startup, person and investor.
- **Operation/Planning** concerns operational matters such as sales, localization, marketing, supply chain, digitization, advisory, company culture and legal regulation.
- **IP and R&D**: IP (intellectual property) and R&D (research and development) can approximate the startups' originality and innovativeness.
- The **customer**, **financial** and **M&A**[10] features are, most of the time, unavailable publicly, which resonates with their scarcity in Figure 4.

3.1 Noticeable Trends

The surveyed literature reflects several trends, summarized below, concerning selecting the input features for DL models.

Single-modal→multi-modal: although the *tabular (aggregated numerical/categorical data)* form still dominates, we see other emerging data modalities: *text*, *graph*, *time-series*, *image*, *video* and *audio*. The relative adoption of different modalities is shown in Figure 4. Especially, a few recent work (e.g., (Shi et al., 2021; Cheng et al., 2019)) has looked into combining multiple input modalities (i.e., multi-modal).

Structured(aggregated)→unstructured(raw): the modalities excluding "tabular" in Figure 4 are unstructured, which become increasingly important as a complement to the structured data (Lyu et al., 2021; Chen et al., 2021; Gastaud et al., 2019), or as a standalone input to the model (Zhang et al., 2021; Tang et al., 2022). Since raw, unstructured data often has a large scale and contains intact-yet-noisy signal, it may bring forward superior performance as long as a proper DL approach is applied (Garkavenko et al., 2022).

Proprietary→paid→free: all data sources utilized in DL-based methods are sorted in Figure 6 according to their occurrences. The traditional proprietary sources are not favored any more due to the limitation of scale and shareability. Paid data sources (e.g., Crunchbase and Pitchbook) are still very popular, because they are mostly quite affordable and well organized. However, neither paid or proprietary data is up-to-date or fine-grained, leading to the increasing adoption of free sources like web page scraping (Garkavenko et al., 2022).

[9]While intrinsic features act from within a company, extrinsic ones wield their influence from the outside. The former often can be controlled by the startup, but the latter can not.

[10]M&A (merger and acquisition) refers to a business transaction in which the ownership of companies (or their operating units) are transferred to or consolidated with another company.

Category	Description of Common Features	The Reference[ref] of Example Work	#ref
Funding	Total number of funding rounds and amount raised	(Allu and Padmanabhuni, 2022; Yin et al., 2021; Horn, 2021; Stahl, 2021) ...	10
	Funding type (e.g., angel and series A/B/C)	(Dellermann et al., 2021; Stahl, 2021; Yeh and Chen, 2020; Sharchilev et al., 2018) ...	8
	Elapsed time since latest funding	(Garkavenko et al., 2022; Ang et al., 2022; Gastaud et al., 2019) ...	6
	Size and type of the latest funding	(Ang et al., 2022; Garkavenko et al., 2022; Ross et al., 2021; Gastaud et al., 2019)	4
	Size and type of seed funding	(Dellermann et al., 2021; Bai and Zhao, 2021; Lyu et al., 2021)	3
	Average per-round statistics	(Garkavenko et al., 2022; Ang et al., 2022; Garkavenko et al., 2021)	3
	Average time between consecutive rounds	(Ross et al., 2021; Garkavenko et al., 2021; Sharchilev et al., 2018)	3
	The raw time-series of funding rounds	(Chen et al., 2021; Stahl, 2021; Horn, 2021)	3
	Accumulated amount for different funding types	(Ross et al., 2021; Sharchilev et al., 2018)	2
	Total amount raised from VC	(Dellermann et al., 2021; Ross et al., 2021)	2
	Post-money valuation of rounds	(Garkavenko et al., 2021)	1
Product/ Service	Industry/sector/sub-sector	(Ang et al., 2022; Ghassemi et al., 2020; Sharchilev et al., 2018; Yu et al., 2018) ...	11
	Textual product description	(Chen et al., 2021; Kim et al., 2020; Cheng et al., 2019; Lee et al., 2018) ...	9
	Project specification on crowdfunding platforms	(Yeh and Chen, 2020; Cheng et al., 2019; Yu et al., 2018; Kim and Park, 2017) ...	7
	Image, video or audio of the product/service	(Tang et al., 2022; Shi et al., 2021; Kaminski and Hopp, 2020; Cheng et al., 2019) ...	5
	Time to market, novelty and differentiation	(Bai and Zhao, 2021; Dellermann et al., 2021; Sharchilev et al., 2018)	3
	Technology maturity, novelty and differentiation	(Allu and Padmanabhuni, 2022; Dellermann et al., 2021; Bai and Zhao, 2021)	3
	Customer focus (e.g., B2B/B2C/B2B2C)*	(Stahl, 2021; Dellermann et al., 2021)	2
	Quality, market penetration and traction	(Bai and Zhao, 2021)	1
	Business models† and scalability	(Dellermann et al., 2021)	1
	The number of product varieties	(Sharchilev et al., 2018)	1
	Textual product review and comment	(Lee et al., 2018)	1
Meta Info.	Founded date and geographical location	(Chen et al., 2021; Garkavenko et al., 2021; Sharchilev et al., 2018; Yu et al., 2018) ...	16
	Has Facebook/Linkedin/Twitter account	(Shi et al., 2021; Dellermann et al., 2021; Ross et al., 2021; Kim and Park, 2017) ...	5
	Domain name or homepage URL	(Ross et al., 2021; Srinivasan et al., 2020; Kim and Park, 2017)	3
	Company legal name and aliases	(Ross et al., 2021; Srinivasan et al., 2020)	2
	Office count and age	(Garkavenko et al., 2022; Sharchilev et al., 2018)	2
	Registered address, email and phone number	(Ross et al., 2021)	1
	Incubator or accelerator support	(Dellermann et al., 2021)	1
Founder Owner	Founding team size (number of co-founders)	(Garkavenko et al., 2021; Ross et al., 2021; Gastaud et al., 2019) ...	11
	Founders' (successful) founding/industry experience	(Bai and Zhao, 2021; Shi et al., 2021; Yeh and Chen, 2020; Srinivasan et al., 2020) ...	11
	Gender, ethnicity or education (uni., major and year)	(Lyu et al., 2021; Ross et al., 2021; Kaiser and Kuhn, 2020; Corea, 2019) ...	8
	Founder ID and score from 3rd-party data sources	(Shi et al., 2021; Yeh and Chen, 2020; Srinivasan et al., 2020; Sharchilev et al., 2018)	4
	Skill (e.g., leadership, sales, law, finance, marketing)	(Bai and Zhao, 2021; Ghassemi et al., 2020; Pasayat et al., 2020; Bento, 2018)	4
	Social capital‡	(Shi et al., 2021; Srinivasan et al., 2020)	2
	Founders' biography (text) and photo	(Srinivasan et al., 2020; Kim and Park, 2017)	2
	Founders' entrepreneurial vision and dedication	(Bai and Zhao, 2021; Dellermann et al., 2021)	2
Team	Team size of all or different functions	(Ang et al., 2022; Garkavenko et al., 2022; Ross et al., 2021; Kim et al., 2020) ...	6
	Completeness and capability of managers and board	(Garkavenko et al., 2021; Bai and Zhao, 2021; Sharchilev et al., 2018)	3
	The time-series of team size	(Stahl, 2021; Horn, 2021)	2
	Statistics of new hire or leavers	(Garkavenko et al., 2021; Sharchilev et al., 2018)	2
	Team composition (e.g., diversity and gender)	(Ross et al., 2021; Sharchilev et al., 2018)	2
	Educational degrees, vocational skill and experience	(Garkavenko et al., 2021; Ross et al., 2021)	2
	3rd-party team score and person ID	(Ghassemi et al., 2020; Sharchilev et al., 2018)	2
	Employees from renowned organizations	(Chen et al., 2021)	1
	Balance/empowerment/competence of the project team	(Yeh and Chen, 2020)	1
Investor	The number of total/distinct investors	(Ferrati et al., 2021; Chen et al., 2021; Kim et al., 2020; Sharchilev et al., 2018) ...	8
	Investor rank by reputation, experience and performance	(Stahl, 2021; Yin et al., 2021; Ferrati et al., 2021; Sharchilev et al., 2018)	4
	VC syndicate (e.g., advantage, diversity and centrality)	(Gastaud et al., 2019; Shin, 2019; Hochberg et al., 2007; Nahata, 2008)	4
	Share and involvement time of each investor	(Sharchilev et al., 2018)	1
Web	Rank/count/duration/bounce rate of website visit	(Garkavenko et al., 2022; Dellermann et al., 2021; Stahl, 2021) ...	5
	The count (aggregated or **timeseries**) of published news	(Yin et al., 2021; Garkavenko et al., 2021; Gastaud et al., 2019; Sharchilev et al., 2018)	4
	Topic or sentiment of **news/articles**	(Garkavenko et al., 2022; Kim et al., 2020; Sharchilev et al., 2018)	3
	Twitter statistics (e.g., followers, tweets and sentiment)	(Garkavenko et al., 2022, 2021; Dellermann et al., 2021)	3
	Count of web pages and domain names	(Garkavenko et al., 2022; Dellermann et al., 2021; Sharchilev et al., 2018)	3
Context	The number of direct competitors	(Allu and Padmanabhuni, 2022; Pasayat and Bhowmick, 2021; Xiang et al., 2012) ...	8
	Funding raised by competitors	(Stahl, 2021; Gastaud et al., 2019)	2
	Per-industry prosperity of the hosting geo-location	(Yin et al., 2021; Gastaud et al., 2019)	2
	Country/state/sector economy and financing env.	(Ross et al., 2021; Yin et al., 2021)	2
	Market/industry size and growth rate	(Allu and Padmanabhuni, 2022)	1
Connection	**The raw company-person-investor graph**	(Allu and Padmanabhuni, 2022; Pasayat and Bhowmick, 2021; Xiang et al., 2012)	3
	Pre-calculated graph features (e.g., betweenness)	(Bonaventura et al., 2020; Liang and Yuan, 2016; Hochberg et al., 2007)	3
Operation/ Planning	Planned revenue model	(Allu and Padmanabhuni, 2022; Dellermann et al., 2021; Bai and Zhao, 2021)	3
	Global exposure and internationalization	(Sharchilev et al., 2018)	1
	Market positioning and go-to-market strategy	(Bai and Zhao, 2021)	1
	Technological surveillance	(Allu and Padmanabhuni, 2022)	1
IP and/ R&D	The number, category and growth of patents	(Kinne and Lenz, 2021; Ferrati et al., 2021; Ross et al., 2021; Kim et al., 2020)	4
	University partnership	(Dellermann et al., 2021)	1
Customer	Customer satisfaction/loyalty	(Chen et al., 2021)	1
	The number of pilot customers	(Dellermann et al., 2021)	1
Financial	Revenue and/or turnover	(Kim et al., 2020; Cao et al., 2022a)	2
M&A	The number of acquisitions	(Ross et al., 2021)	1
Data	The total number of events/records	(Kim et al., 2020)	1

* Common types of customer focus: B2B: business-to-business. B2C: business-to-consumer. B2B2C: business-to-business-to-consumer, where businesses access customers via a 3rd-party.
† Business models include many, such as subscription centric, freemium, cross selling, hidden revenue, no frills, and layer player.
‡ Social capital is a positive product of human interactions, which comprises two aspects: bonding (intra group) and bridging (inter groups). Nowadays, it is increasingly represented by activities on social media and applications (Shi et al., 2021).

Table 1: The feature categories and the commonly adopted features within each categy. Due to limited space, we can not list all publications that adopt the corresponding feature, but the right-most "#ref" column indicates the total number of occurrences for each feature. Most of the features are structured numerical/categorical input, and we use **boldface** to emphasize the unstructured features.

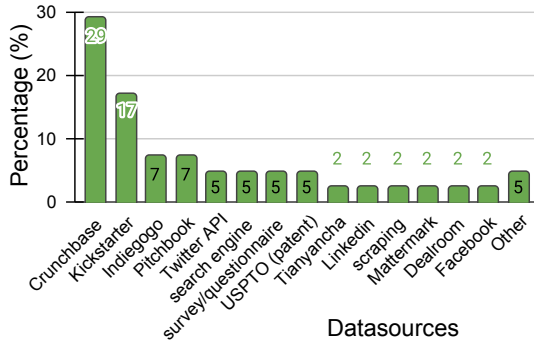


Figure 6: The occurrences of common data sources: **paid** sources are Crunchbase, Pitchbook, Tianyancha, Linkedin, Mattermark, Dealroom; **free** sources are Kickstarter/Indiegogo/scraping, Twitter API, search engines (e.g., Google), USPTO (United States Patent and Trademark Office, Facebook (the pages about startups); **proprietary** data are usually only accessible from investment firms (in "Other" category), governmental/administrative departments or survey/questionnaire.

About dataset size: to understand how many samples (the number of companies) researchers use for training their DL models, we plot the distribution/histogram of dataset size in Figure 8. It shows a median and average size of 35,621 and 107,694 respectively, which is expected to continue to grow.

Intrinsic(independent)→extrinsic(contextual): classically, most factors driving investors' decisions would be only *independent and intrinsic*[9] to the startup, most notably at the expense of *extrinsic and contextualized*[9] features (Gastaud et al., 2019). The community has started steering towards using more context and connection features.

4 Data Split

Splitting the dataset is a mandatory step before training any ML/DL model, yet it is often discussed very lightly (sometimes even neglected) in the literature on startup success prediction. It is generally recommended to divide the dataset into non-overlapping *training* ($\mathbf{x}_{\text{train}}$), *evaluation* ($\mathbf{x}_{\text{eval}}$) and *test* ($\mathbf{x}_{\text{test}}$) subsets. The model will be trained solely on $\mathbf{x}_{\text{train}}$. Hyper-parameters are searched using $\mathbf{x}_{\text{eval}}$. In the simplest form, the training will be run for N times with different hyper-parameters, resulting in N trained models, each of which is evaluated on $\mathbf{x}_{\text{eval}}$. The best performing model on $\mathbf{x}_{\text{eval}}$ should be tested on $\mathbf{x}_{\text{test}}$ before deployment.

4.1 Company-Centric vs. Investor-Centric

To predict the success of startups, the appropriate way to split the dataset is not as straightforward as it appears in ML/DL researches for other domains. We visualize a minimal example in Figure 7 to facilitate our discussion; there are three startups (A, B and C) founded at different dates over the timeline. According to some success criteria (Section 2), A and B are labeled as positive (i.e., promising investing targets: $y^{(\text{A})}=y^{(\text{B})}=1$) some time after they are founded. The majority become unfavourable (e.g., the label of C is $y^{(\text{C})}=0$) to VC, if no sign of success some years after their founding dates.

With a **company-centric** view, one can choose some event types (e.g., seed and pre-A rounds), the dates of which are *feature snapshot dates*. We can then compute one sample using data before each snapshot date. As shown in Figure 7, there are three snapshot dates on the timeline of startup A, leading to three samples (i.e., $\mathbf{x}_1^{(\text{A})}$, $\mathbf{x}_2^{(\text{A})}$ and $\mathbf{x}_3^{(\text{A})}$) that are all labeled positive (i.e., $y_1^{(\text{A})}=y_2^{(\text{A})}=y_3^{(\text{A})}=1$). In a sense, startup A is augmented by generating three $\langle$sample, label$\rangle$ pairs: $\langle \mathbf{x}_1^{(\text{A})}, y_1^{(\text{A})} \rangle$, $\langle \mathbf{x}_2^{(\text{A})}, y_2^{(\text{A})} \rangle$ and $\langle \mathbf{x}_3^{(\text{A})}, y_3^{(\text{A})} \rangle$. Similarly, B and C create another four pairs: $\langle \mathbf{x}_1^{(\text{B})}, y_1^{(\text{B})} \rangle$, $\langle \mathbf{x}_2^{(\text{B})}, y_2^{(\text{B})} \rangle$, $\langle \mathbf{x}_1^{(\text{C})}, y_1^{(\text{C})} \rangle$ and $\langle \mathbf{x}_2^{(\text{C})}, y_2^{(\text{C})} \rangle$. The company-centric split will randomly allocate these pairs into one of the sets (training, evaluation or test), as in work such as (Ang et al., 2022; Yeh and Chen, 2020).

With an **investor-centric** view, as in work like (Wu et al., 2022; Ferrati et al., 2021), the feature snapshot dates are randomly sampled (before the corresponding label date), therefore they do not represent any event(s). More importantly, the global timeline is fragmented (from earliest startup founding date to now) into three periods, i.e., training, evaluation and test period, as illustrated in Figure 7. For a startup, the period that its label belongs determines the dataset split it should go to. Applying this rule, we can see (cf. Figure 7) that the three $\langle$sample, label$\rangle$ pairs from A should go to the training set; the two pairs from B belong to the test set; and lastly, the two pairs from C will head to the evaluation set. (Sharchilev et al., 2018) claims that **investor-centric view is preferred**, since it better resembles the real-world scenario of how investment professionals predict the success of startups.

4.2 Data Generation Process Matters

When assembling the samples (i.e., $\mathbf{x}_i^{(\cdot)}$ in Figure 7) using data up till the snapshot dates, one should make sure that no future information is leaked into $\mathbf{x}_i^{(\cdot)}$. This requires in-depth understanding of not only the data itself (*know-what*) but also the data generation process (*know-how*), which we found

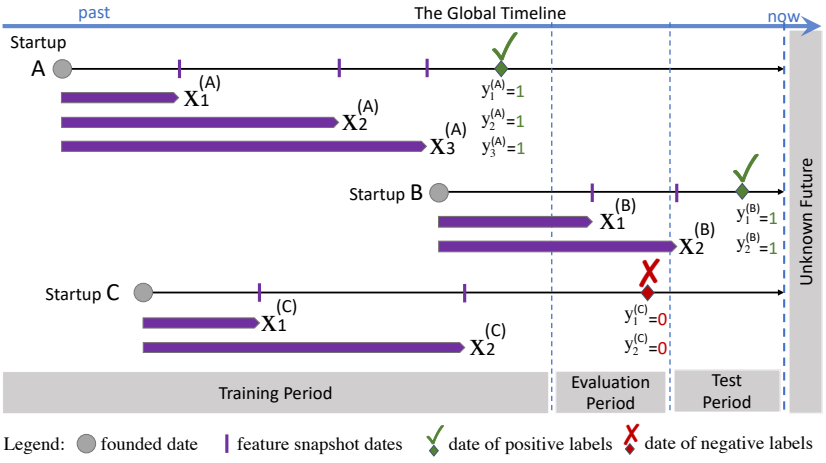


Figure 7: Visualization of investor-centric split using three example startups.

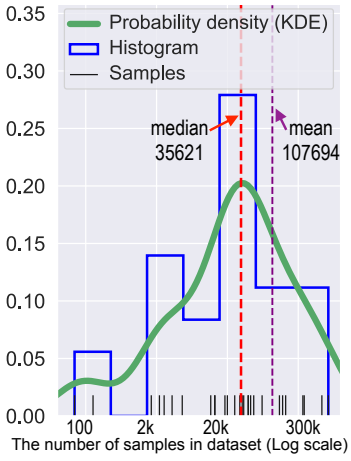


Figure 8: Dataset size distribution.

is seldomly addressed by the literature. We hereby give a concrete example out of many: a startup in the dataset has an annual revenue data point (from BvD[11]) with a timestamp 2020-12-31; but this data point should be ignored when predicting on 2021-06-01. The reason is that fiscal reports (the source of revenue data) often have a delay of about 12 months, causing the 2020-12-31 data point unavailable until (earliest) 2021-12-31. Without examining such matters, the model performance in production may fail catastrophically.

5 Evaluation Strategy

The decision of deploying any model is often made by looking at the evaluation results. To achieve that, some *evaluation metrics* are employed to measure the quality of predictions y by comparing to the ground-truth labels $\hat{y}$. The metric values computed over the evaluation set (i.e., $\mathbf{x}_{\text{eval}}$) are used to determine which model (among many trained using different hyper-parameters) will be deployed for production eventually. This process also fulfills the objective of hyper-parameter search. It has been discussed in Section 4 that the evaluation metrics should also be calculated on the test set $\mathbf{x}_{\text{test}}$ as an indication of the model's generalization capability.

The evaluation metrics adopted in the DL literature include (ordered by their occurrences as shown in Figure 9 with an example citation) *precision* (Zhang et al., 2021), *ROC-AUC* (*area under the receiver operating characteristics*) (Ross et al., 2021), *accuracy* (Bai and Zhao, 2021), *FPR* (*false-positive rate*) (Ghassemi et al., 2020), *TPR* (*true-positive rate*) (Garkavenko et al., 2022), *hit rate* (Allu and Padmanabhuni, 2022), *NDCG* (*normalized discounted cumulative gain*) (Chen et al., 2021), *portfolio simulation* (Yin et al., 2021), *RMSE* (*root mean square deviation*) (Wu et al., 2022), *AUPR* (*area under the precision-recall curve*) (Zhang et al., 2021), *average precision* (Lyu et al., 2021), *confusion matrix* (Ross et al., 2021), *F0.1 score* (Sharchilev et al., 2018), *MAE* (*mean absolute error*) (Wu et al., 2022), *MCC* (*Matthews correlation coefficient*) (Dellermann et al., 2021), *PR* (*precision-recall*) *curve* (Stahl, 2021), and R^2 (Garkavenko et al., 2021).

Most trained models are expected to serve as a decision-support system for VC deal sourcing. Realistically, human professionals are only able to assess a limited amount of startups. Further, because of fund size limitation, investors can only fund a very small fraction of startups (Stahl, 2021). As a result, the **evaluation metric should aim for high-precision** (corresponding to high-certainty and low-recall)[12] (Sharchilev et al., 2018), which explains the popularity of *precision*, *TPR*, *FPR*, *hit rate* and *F0.1 score* in Figure 9.

5.1 Portfolio Simulation

There are four key questions to answer concerning any model trained to facilitate VC deal sourcing: **Q1** What is the expected success ratio (or ROI) of the portfolio (with different sizes) constructed according to model predictions? **Q2** How will the model-driven portfolio perform in relation to the historical records of renowned investment firms? **Q3** Is the model significantly superior than a random policy? **Q4** How far does the model fall be-

[11] Bureau van Dijk: www.bvdinfo.com

[12] In the scope of VC deal sourcing, high-precision means the rate of "correct" prediction within the top-N list (i.e., TPR) should be high. According to the typical PR curve, precision tends to be higher for smaller N; yet recall suffers from it.

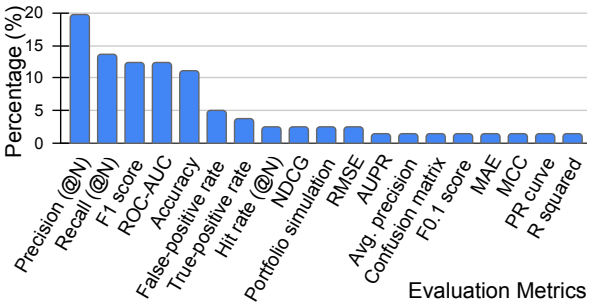


Figure 9: The distribution of adopted evaluation metrics. The notation "@N" implies the corresponding metric is calculated over a top-N list. Precision, hit rate, and F0.1 score are popular metrics with a focus of high-precision. Portfolio simulation suited particularly well to startup success prediction, while others are general-purpose metrics for evaluating ML/DL models.

hind a theoretical perfect portfolio with 100% success ratio? Answering all questions simultaneously using any single general-purpose ML/DL metric is challenging and sometimes far-fetched. To that end, some recent works (Ross et al., 2021; Yin et al., 2021) (though still far from a wide adoption according to Figure 9) have emerged proposing to **evaluate via portfolio simulations**. Recall that in Section 4, we recommended the investor-centric dataset split demonstrated in Figure 7. With that split, we make the trained models to predict the conditional success probability of each startup in evaluation/test subset, using the end date of training period as the feature snapshot date. Then, we construct an investment portfolio of size k by selecting top-k startups with the highest predicted probabilities. As an indication of portfolio performance, we count the number of startups that eventually obtain a positive label. The portfolio size k should be varied, so that we can plot one performance curve (the four colored curves in Figure 10) for each model. To answer **Q1**, a steeper curve corresponds to a better model. The performance of a perfect model is a diagonal line, implying all portfolio startups will succeed. To address **Q2**, one just needs to measure the angular distance to diagonal. The simplest possible model is a random policy, the performance of which is represented by the flattest straight-dashed line in Figure 10; the angular distance between this "random" line to any model's curve answers **Q3**. Finally, the historical fund performance of investment firms can be easily plotted as individual points, the vertical distances from which to models' curves give insights for **Q4**. In practice, the investment firms are more constrained than simulation: they can not invest in any startup due to many

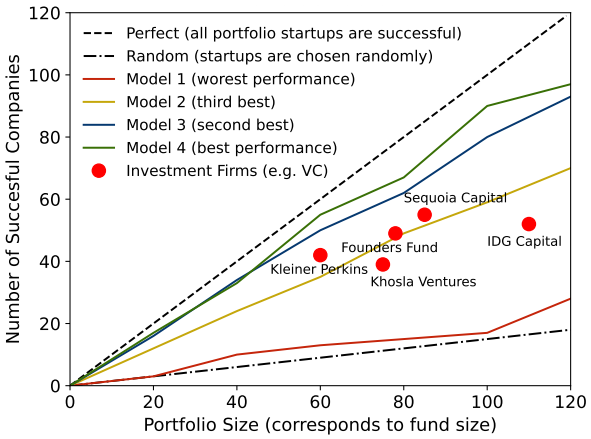


Figure 10: Portfolio simulation. The trained DL model is used to form portfolios of size $k \in \{20, 40, \ldots, 120\}$ (x-axis); the number of eventually successful startups is plotted against the corresponding k, resulting in a performance curve (cf. the colored curves). The perfect/random cases (dashed lines) and performance of investment firms (red dots) can be plotted as well for comparison. It is adapted from (Halvardsson, 2023).

reasons like founders preference, portfolio conflict and investment mandate. This constraint becomes more prominent when investors compete to invest in startups with great success potential.

6 Conclusion

Finding the rare unicorn startups is a challenging task, hence often regarded as the holy grail for early-stage investors like Venture Capital firms. To avoid entirely relying on human domain expertise and intuition, investors usually employ data-driven approaches to forecast the success probability of startups. The rapid growth of data volume and variety makes deep learning (DL) a potentially superior approach to address this task. To the best of our knowledge till this date, there has not been any comprehensive survey on this topic. According to our synthesis of carefully selected literature, the innovation mostly lies in how an off-the-shelf DL model is applied, rather than in the model itself. So we focus on summarizing our understanding and learning concerning four key aspects:

- Optimization target: consider a mixture of criteria, while prioritizing the short-term event.
- Feature selection: scale the dataset with multimodal, unstructured, free and extrinsic features.
- Data split: apply the investor-centric split with the knowledge of data generation process.
- Evaluation strategy: pick the metrics aiming for high-precision and perform portfolio simulation.

Finally, authors' outlook of DL adoption in startup success prediction is three fold: (1) more easy-to-use software tools will be developed to promote good practices and lower the barrier to entry; (2) the majority of the available data is unlabeled and small scaled, hence more data/label efficient DL models will be proposed; (3) data privacy and model security will gain more emphasis in the coming years.

Acknowledgments

We are especially grateful to Alex Patow, Anton Ask Åström, Armin Catovic, Ashley Lundström, Dhiana Deva, Drew McCornack, Richard Stahl and Sofie Grant from EQT Group for the discussion and feedback on a more comprehensive and detailed version of this study (Cao et al., 2022b). We also thank the interest and input from Celine Xu (H&M×AI & McKinsey), Daniel Wroblewski (CPPIB), Fenni Kang (AntAlpha & Barclays), Gustaf Halvardsson (KTH & EQT), Rockie Yang (Knock Data) and Wenbing Huang (Tsinghua Uni.). This work is also generally supported by the entire EQT Motherbrain team.

References

Zoltan J Acs and Laszlo Szerb. 2007. Entrepreneurship, economic growth and public policy. *Small business economics*, 28(2):109–122.

Ramakrishna Allu and Venkata Nageswara Rao Padmanabhuni. 2022. Predicting the success rate of a start-up using lstm with a swish activation function. *Journal of Control and Decision*, 9(3):355–363.

Yu Qian Ang, Andrew Chia, and Soroush Saghafian. 2022. Using machine learning to demystify startups' funding, post-money valuation, and success. In *Innovative Technology at the Interface of Finance and Operations*, pages 271–296. Springer.

John Baer and Sharon S McKool. 2014. The gold standard for assessing creativity. *International Journal of Quality Assurance in Engineering and Technology Education (IJQAETE)*, 3(1):81–93.

Sarah Bai and Yijun Zhao. 2021. Startup investment decision support: Application of venture capital scorecards using machine learning approaches. *Systems*, 9(3):55.

Francisco Ramadas da Silva Ribeiro Bento. 2018. *Predicting start-up success with machine learning*. Ph.D. thesis, Universidade NOVA de Lisboa.

Steve Blank. 2013. Why the lean start-up changes everything. *Harvard business review*, 91(5):63–72.

Moreno Bonaventura, Valerio Ciotti, Pietro Panzarasa, Silvia Liverani, Lucas Lacasa, and Vito Latora. 2020. Predicting success in the worldwide start-up network. *Scientific Reports*, 10(1):1–6.

Lele Cao, Sonja Horn, Vilhelm von Ehrenheim, Richard Anselmo Stahl, and Henrik Landgren. 2022a. Simulation-informed revenue extrapolation with confidence estimate for scaleup companies using scarce time series data. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA*, page 12 pages, New York, NY, USA. Association for Computing Machinery (ACM).

Lele Cao, Vilhelm von Ehrenheim, Sebastian Krakowski, Xiaoxue Li, and Alexandra Lutz. 2022b. Using deep learning to find the next unicorn: A practical synthesis. *arXiv preprint arXiv:2210.14195*.

Miao Chen, Chao Wang, Chuan Qin, Tong Xu, Jianhui Ma, Enhong Chen, and Hui Xiong. 2021. A trend-aware investment target recommendation system with heterogeneous graph. In *Intl. Joint Conf. on Neural Networks*, pages 1–8.

Chaoran Cheng, Fei Tan, Xiurui Hou, and Zhi Wei. 2019. Success prediction on crowdfunding with multimodal deep learning. In *International Joint Conference on Artificial Intelligence*, pages 2158–2164.

Sergey Chernenko, Josh Lerner, and Yao Zeng. 2021. Mutual funds as venture capitalists? evidence from unicorns. *The Review of Financial Studies*, 34(5):2362–2410.

Francesco Corea. 2019. AI and venture capital. In *An introduction to data*, pages 101–110. Springer.

Douglas Cumming and Na Dai. 2010. Local bias in venture capital investments. *Journal of empirical finance*, 17(3):362–380.

Dominik Dellermann, Nikolaus Lipusch, Philipp Ebel, Karl Michael Popp, and Jan Marco Leimeister. 2021. Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method. In *International Conference on Information Systems*.

Francesco Ferrati, Haiquan Chen, and Moreno Muffatto. 2021. A deep learning model for startups evaluation using time series analysis. In *European Conf. on Innovation and Entrepreneurship*, page 311. Academic Conferences limited.

Mariia Garkavenko, Eric Gaussier, Hamid Mirisaee, Cédric Lagnier, and Agnès Guerraz. 2022. Where do you want to invest? predicting startup funding from freely, publicly available web info. *arXiv preprint arXiv:2204.06479*.

Mariia Garkavenko, Hamid Mirisaee, Eric Gaussier, Agnès Guerraz, and Cédric Lagnier. 2021. Valuation of startups: A machine learning perspective. In *European Conference on Information Retrieval*, pages 176–189. Springer.

Clement Gastaud, Theophile Carniel, and Jean-Michel Dalle. 2019. The varying importance of extrinsic factors in the success of startup fundraising: competition at early-stage and networks at growth-stage. *arXiv preprint arXiv:1906.03210*.

M Ghassemi, C Song, and T Alhanai. 2020. The automated venture capitalist: Data and methods to predict the fate of startup ventures. In *AAAI Workshop on Knowledge Discovery from Unstructured Data in Financial Services*.

Paul A Gompers, Will Gornall, Steven N Kaplan, and Ilya A Strebulaev. 2020. How do venture capitalists make decisions? *Journal of Financial Economics*, 135(1):169–190.

Gustaf Halvardsson. 2023. A transformer-based scoring approach for startup success prediction. Master's thesis, KTH Royal Institute of Technology & EQT Partners.

Yael V Hochberg, Alexander Ljungqvist, and Yang Lu. 2007. Whom you know matters: Venture capital networks and investment performance. *The Journal of Finance*, 62(1):251–301.

Sonja Horn. 2021. Deep learning models as decision support in venture capital investments: Temporal representations in employee growth forecasting of startup companies. Master's thesis, KTH Royal Institute of Technology & EQT Partners.

Ari Hyytinen, Mika Pajarinen, and Petri Rouvinen. 2015. Does innovativeness reduce startup survival rates? *Journal of business venturing*, 30(4):564–581.

Ulrich Kaiser and Johan M Kuhn. 2020. The value of publicly available, textual and non-textual information for startup performance prediction. *Journal of Business Venturing Insights*, 14:e00179.

Jermain C Kaminski and Christian Hopp. 2020. Predicting outcomes in crowdfunding campaigns with textual, visual, and linguistic signals. *Small Business Economics*, 55(3):627–649.

Hyoung J Kim, Tae San Kim, and So Y Sohn. 2020. Recommendation of startups as technology cooperation candidates from the perspectives of similarity and potential: A deep learning approach. *Decision Support Systems*, 130:113229.

Jongho Kim and Jiyong Park. 2017. Does facial expression matter even online? an empirical analysis of facial expression of emotion and crowdfunding success. In *International Conference on Information Systems*.

Jan Kinne and David Lenz. 2021. Predicting innovative firms using web mining and deep learning. *PloS One*, 16(4):e0249071.

SeungHun Lee, KangHee Lee, and Hyun-chul Kim. 2018. Content-based success prediction of crowdfunding campaigns: A deep learning approach. In *Companion of the ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 193–196.

Yuxian E Liang and Soe-Tsyr D Yuan. 2016. Predicting investor funding behavior using Crunchbase social network features. *Internet Research: Electronic Networking Applications and Policy*, 26(1):74–100.

Robert N Lussier and Claudia E Halabi. 2010. A three-country comparison of the business success versus failure prediction model. *Journal of Small Business Management*, 48(3):360–377.

Robert N Lussier and Sanja Pfeifer. 2001. A crossnational prediction model for business success. *Journal of Small Business Management*, 39(3):228–239.

Shiwei Lyu, Shuai Ling, Kaihao Guo, Haipeng Zhang, Kunpeng Zhang, Suting Hong, Qing Ke, and Jinjie Gu. 2021. Graph neural network based VC investment success prediction. *arXiv preprint arXiv:2105.11537*.

Max Marmer, Bjoern L Herrmann, Ertan Dogrultan, Ron Berman, Cuck Eesley, and Steve Blank. 2011. Startup genome report extra: premature scaling. *Startup Genome*, 10:1–56.

Rajarishi Nahata. 2008. Venture capital reputation and investment performance. *Journal of Financial Econ.*, 90(2):127–151.

Ajit Kumar Pasayat and Bhaskar Bhowmick. 2021. An evolutionary algorithm-based framework for determining crucial features contributing to the success of a start-up. In *IEEE Technology and Engineering Management Conference-Europe (TEMSCON-EUR)*, pages 1–6. IEEE.

Ajit Kumar Pasayat, Bhaskar Bhowmick, and Ritik Roy. 2020. Factors responsible for the success of a start-up: A meta-analytic approach. *IEEE Transactions on Engineering Management*.

Greg Ross, Sanjiv Das, Daniel Sciro, and Hussain Raza. 2021. CapitalVX: A machine learning model for startup selection and exit prediction. *The Journal of Finance and Data Science*, 7:94–114.

José Santisteban, David Mauricio, Orestes Cachay, et al. 2021. Critical success factors for technology-based startups. *International Journal of Entrepreneurship and Small Business*, 42(4):397–421.

Scott Shane. 2012. The importance of angel investing in financing the growth of entrepreneurial ventures. *The Quarterly Journal of Finance*, 2(02):1250009.

Boris Sharchilev, Michael Roizner, Andrey Rumyantsev, Denis Ozornin, Pavel Serdyukov, and Maarten de Rijke. 2018. Web-based startup success prediction. In *International Conference on Information and Knowledge Management*, pages 2283–2291.

Jiatong Shi, Kunlin Yang, Wei Xu, and Mingming Wang. 2021. Leveraging deep learning with audio analytics to predict the success of crowdfunding projects. *The Journal of Supercomputing*, 77(7):7833–7853.

Sang Yoon Shin. 2019. Network advantage's effect on exit performance: examining venture capital's inter-organizational networks. *International Entrepreneurship and Management Journal*, 15(1):21–42.

Eulalia Skawińska and Romuald I Zalewski. 2020. Success factors of startups in the EU - a comparative study. *Sustainability*, 12(19):8200.

Arvind Srinivasan et al. 2020. An ensemble deep learning approach to explore the impact of enticement, engagement and experience in reward based crowdfunding. Working paper, Department of Computer Science and Engineering, SRM Institute of Science and Technology.

Richard Hermann Anselmo Stahl. 2021. Leveraging time-series signals for multi-stage startup success prediction. Master's thesis, ETH Zurich & EQT Partners.

Zhe Tang, Yi Yang, Wen Li, Defu Lian, and Lixin Duan. 2022. Deep cross-attention network for crowdfunding success prediction. *IEEE Transactions on Multimedia*.

David Teten, Adham AbdelFattah, Koen Bremer, and Gyorgy Buslig. 2013. The lower-risk startup: how venture vapitalists increase the odds of startup success. *The Journal of Private Equity*, 16(2):7–19.

Cemre Ünal and Ioana Ceasu. 2019. A machine learning approach towards startup success prediction. IRTG 1792 Discussion Paper 2019-022, Berlin.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Likang Wu, Zhi Li, Hongke Zhao, Qi Liu, and Enhong Chen. 2022. Estimating fund-raising performance for start-up projects from a market graph perspective. *Pattern Recognition*, 121:108204.

Guang Xiang, Zeyu Zheng, Miaomiao Wen, Jason Hong, Carolyn Rose, and Chao Liu. 2012. A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on TechCrunch. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 607–610.

Jen-Yin Yeh and Chi-Hua Chen. 2020. A machine learning approach to predict the success of crowdfunding fintech project. *Journal of Enterprise Information Management*.

Dafei Yin, Jing Li, and Gaosheng Wu. 2021. Solving the data sparsity problem in predicting the success of the startups with machine learning methods. *arXiv preprint arXiv:2112.07985*.

Pi-Fen Yu, Fu-Ming Huang, Chuan Yang, Yu-Hsin Liu, Zi-Yi Li, and Cheng-Hung Tsai. 2018. Prediction of crowdfunding project success with deep learning. In *International Conference on E-Business Engineering*, pages 1–8. IEEE.

Shengming Zhang, Hao Zhong, Zixuan Yuan, and Hui Xiong. 2021. Scalable heterogeneous graph neural networks for predicting high-potential early-stage startups. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2202–2211.