POSE: Phase II: RedPoint: A Sustainable Open Source Ecosystem for Industry-Academia Collaboration Towards Intelligent Time Series Analysis.

Proposal submission on Research.gov

.[DEADLINE: October 21, 2022.]

- ☐ Project Summary [One (1) page max].
- ☐ Documentation
  - ☐ Collaborators and other affiliations
  - ☐ Current and pending support
  - ☐ Bio sketch
  - ☐ Budget
  - ☐ Data management
  - ☐ Equipment and facilities
  - ☐ Sub-award documentation

Project Description. - (15) pages max-

- ☐ Context of OSE
- ☐ Broader Impacts
- ☐ Ecosystem Growth
- ☐ Organization and Governance
- ☐ Community Building

Letters of collaboration. Minimum 3 and maximum 5.

- ☐ Letter of collaboration from Samsung.
- ☐ Letter of collaboration from CMU.
- ☐ Letter of collaboration from OSU.
- ☐ Letter of collaboration from Angela. Rice, Analog Devices.

Other

- ☐ Security Plan

Project Summary

1 Overview

RedPoint, the proposed open-source ecosystem aims to increase the reach of open-source alliance (Time-series Outlier Detection (TODS), TDengine, B-E-D, and PyOD) which have gathered more than 25,000 stars on GitHub and more than 10 million downloads. RedPoint is designed to help collaborating universities partner with participating companies to tackle distinct business intelligence challenges while enabling human interpretation to help subject matter experts understand, trust, and leverage advanced machine learning (ML) technologies. In addition, the proposed ecosystem intends to grow the current open-source developer community and reach new domain experts through a series of events designed to enable new pathways for the development of collaborative open-source in the time-series domain that could lead to new technology products or services that have broad societal impacts. This POSE phase 2 proposal brings together world-class US universities (Ohio State University, CMU, Texas A&M, and The University of Florida), top corporations that are deploying machine learning at a large scale (Samsung and TAOS data), and AI POW LLC, a small businesses with strong research capabilities.

2 Intellectual Merit

The open-source presented as a foundation of RedPoint, the Open-Source Ecosystem, involves novel and significant advances in ML Automation and Interpretation applied to anomaly detection. Each advance has dramatically extended current ML techniques toward dealing with deep learning challenges raised in real-world business intelligence data. A flexible and modular end-to-end anomaly detection system, different from current empirical approaches, is needed by the open-source community to automatically decide the optimal configuration of complex modeling pipelines to tackle their academic and industrial time-series data challenges.

3 Broader Impacts

The proposed project's successful outcome will expand the impact of AI POW's open-source developments in dealing with an emerging and critical ML automation and interpretation problem with significant industrial applications. One example is that this project's results will have an immediate and substantial impact on improving the manufacturing industry's performance, enabling production line managers to simplify the anomaly identification and prediction task while understanding the flaws' origin to elevate product quality.

Similarly, as data science and engineering become more critical for businesses competing in a globalized digital economy, it is paramount to consolidate a 21st-century data-capable workforce. Many large enterprises can implement analysis detection systems because they can afford a highly specialized data scientist team. The proposed project will help create an open-source ecosystem for domain experts to benefit from affordable and easy-to-use business intelligence digital tools tailored to software engineers with more basic programming understanding. A phase 2 follow-up proposal of this open-source ecosystem can allow a larger group of small and medium businesses to adopt and benefit from advanced ML tools and embrace the data revolution by using their data assets in new innovative ways.

Keywords: CISE; Outlier Detection; Machine Learning; time-series; AutoML.

POSE: Phase II: RedPoint: A Sustainable Open Source Ecosystem for Industry-Academia Collaboration Towards Intelligent Time Series Analysis

1 Introduction

Time-series data is one of the most popular data types used by virtually every industry to analyze trends, plan for the future, and make data-driven business decisions. For example, retailers use time-series data to track product sales and identify opportunities for product development; financial institutions use time-series analysis to detect abnormal customer behaviors to prevent fraud; and HVAC manufacturers use it to monitor such assets remotely and detect anomalies that could lead to quality issues. Time-series data is often considered a goldmine of information because it can reveal patterns that could be used to predict future events. According to Markets and Markets and IBISWorld [1, 2] only the remote asset monitoring market size is expected to grow from USD 16.5 billion in 2020 to USD 32.6 billion by 2024.

However, at the same time, **time-series data is extremely complex and hard to analyze.** The number of variables can be enormous, and their interactions are often not well understood. To make matters worse, time-series data is often noisy because it can be subject to many sources of variation such as measurement errors and random events. This means that organizations may miss many key opportunities due to the difficulty of analyzing time-series data. Consequently, there is a critical need for time-series analysis open-source ecosystem to help organizations (private and public) discover hidden patterns in their data and make better decisions.

Given the importance of time-series data, our team and collaborators developed multiple time-series analysis tools for different use cases and made them available on GitHub over the past few years. The tools are designed to be simple and easy-to-use, yet powerful enough for users who want to gain insights from their time-series data, for example, **TODS**[1], a full-stack automated machine learning system for outlier detection on multivariate time-series data. Our team open-sources all the modules in TODS to provide exhaustive for building machine learning-based outlier detection systems including: data processing, time series processing, feature analysis (extraction), detection algorithms, and reinforcement module; our collaborators at **TDengine** [2] developed an open-source, high-performance,

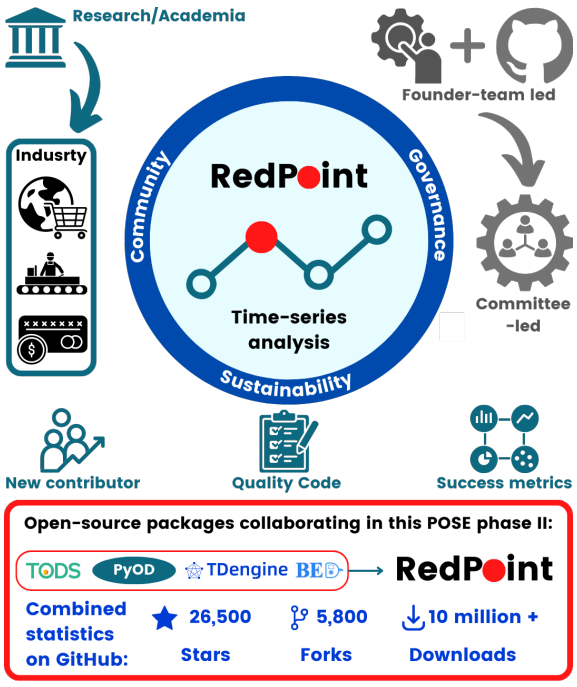

Figure 1: (Top) The three key activities to grow into the mature ecosystem RedPoint, are: (1) to mature our current founder-team-led **governance** model into a board-led model to govern various aspects of RedPoint; (2) a **sustainable** model for developing high quality open-source code and providing onboarding mechanisms so new contributors can easily join RedPoint; and (3) we plan to set periodic **community**-building events to foster collaboration between industry partners and academia collaborators. (Bottom) List of **open-source alliance** projects for the RedPoint ecosystem and GitHub statistics.

[1] github.com/datamllab/tods
[2] github.com/taosdata/TDengine

cloud native time-series database optimized for Internet of Things (IoT), Connected Cars, and Industrial IoT. It enables efficient, real-time data ingestion, processing, and monitoring large scale data per day, generated by billions of sensors and data collectors; our partners at Carnegie Mellon University developed **PyOD**[3], the most comprehensive and scalable open-source Python library for detecting outlying objects in multivariate data. Since 2017, PyOD has been successfully used in numerous academic research and commercial products, with more than 8 million downloads; finally, **B-E-D**[4], end-to-end system designed to make ML models designed for cloud applications to be deployed on Edge Devices. B-E-D, technologies in model compression and hardware aware AutoML solution is being improved to tackle the challenge of anomaly detection on IoT devices. B-E-D will help RedPoint users reduce the possibility of inaccuracies that might arise due to bias or human error.

Many data collected in industry are naturally time-ordered [3,4], such as financial data [5,6], sensor data [7,8], and ECG data [9,10]. For our RedPoint collaborators **Samsung, TAOS data LLC, American Innovations, and AI POW LLC**, analyzing the time-series is an important step to understand the internal structure, trend, or seasonality of the data [11]. Here, we provide a few important downstream applications of time-series that are key to industry:

- **Time-series forecasting:** it is the task of predicting future values given the previous values in a time [12–14]. This task can help organizations anticipate upcoming events. For example, electricity companies can leverage time-series forecasting to predict future electronic usage so that they can make better preparations [15–17].

- **Time-series classification:** it is the task of classifying time-series into certain categories [18–20]. For example, based on the ECG data collected from patients, time-series classification can identify the patients that may be susceptible to diseases [21–23].

- **Time-series outlier detection:** it is the task of identifying the minority patterns in the time-series that significantly deviate from the majorities [24]. Outlier detection on time-series data has various applications, such as fraud detection [25–27], manufacturing fault detection [28–30], healthcare monitoring [31–33].

We want to highlight that our fellow collaborators at Samsung, and Texas A&M University have provided further detail on the importance of intelligent time-series analysis in their letters of support.

The open-source ecosystem we propose, RedPoint, brings together the four communities to offer a flexible and modular end-to-end anomaly detection suite of open-source packages for advanced time-series analysis. **Our goal is to amplify our impact by providing a sustainable open-source ecosystem suitable for industry (with the help of our collaborators at American Innovations and TAOS data) and academia (through our partners at Carnegie Mellon University, Rice, Ohio State and Texas A&M University) to collaborate on new research and also provide a practical solution for real-world applications.** This proposed Pathways to Enable Open-Source Ecosystems (POSE) aims to attract new contributors at every level of technical expertise; from those who want to help us build, test, and refine the software all the way up to those who want to contribute code or data analysis techniques. RedPoint has the potential to consolidate and expand the reach of our open-source ecosystem and allow domain experts with basic programming knowledge to detect, predict, and understand anomalies for their business analysis challenges in ways that previously were extremely difficult, if not impossible. It is important to highlight that

[3] github.com/yzhao062/pyod
[4] github.com/datamllab/BED_main

combined, the four time-series focused communities have attracted more than 26k stars, 5k forks and over 10 million downloads in GitHub.

As shown in Fig. 1, **the three of the key activities to perform with the work presented here are (1) to mature our current founder-team-led governance model** in which the group that started the project also administers the project, establishes its vision, and controls permissions to merge code into it, **into a structured board-led model to govern various aspects of the RedPoint** including new software releases, module management, site administration, and community growth; **(2) a sustainable model for managing and developing open-source code**, securing its quality and accuracy, protecting the privacy of users' information, **and providing onboarding mechanisms so new contributors can easily join the effort**; and **(3)** because building a community around an open-source project is vital, **we plan to set periodic community outreach events, which include in-person conferences, online hackathons, and hybrid-format idea labs which will help foster a sense of community around RedPoint**. We also plan to set up a RedPoint-specific website that will serve as a central hub for users and developers to access information of RedPoint, its features and capabilities, and how to get involved with the project itself or with our local communities.

2 Context of OSE

This POSE phase II proposal brings together the open-source communities: TODS, TDengine, PyOD, B-E-D; three US universities with strong state-of-the-art machine learning research: The Ohio State University, Carnegie Mellon University, The University of Florida, and Texas A&M University; corporations that are deploying machine learning at a large scale: Samsung, American Innovations and TAOS data, and AI POW LLC, a small business with strong research capabilities. This proposal aims to create a community that can provide the following benefits: a place for academic researchers to work on real-world problems and get feedback from industry practitioners while also being a place for industry practitioners to have their problems solved by academics with state-of-the-art techniques, tools, and skills.

The resulting OSE will lead to an open-source community that can easily help our current and new users to access features for their time-series analysis and applications. One immediate goal is to tailor this ecosystem to be a platform that can easily integrate multiple vertical applications depending on the interest of the domain experts. A few verticals where PyOD, TDengine, and TODS have been used are anomaly detection (energy) [34–36], weather forecast (agriculture) [37–39], quality assurance (manufacturing) [40–42], fraud detection (finance) [26, 43, 44] and remote asset monitoring (IoT) [45–47], to name a few.

Our conceptualization of the project vision and key activities in the context of the open-source products and proposed RedPoint activities for this Phase 2 POSE proposal are shown in Fig. 2.

2.1 Existing Publicly-available Open-source Products

The following publicly-available open-source projects have been developed for the data storage, data processing, and downstream applications of time-series. Data storage: **TDEngine**[5] is an open-source, cloud-native database optimized for time-series. It supports high performance querying and data compression. **PyOD**[6] is a unified Python library for outlier detection. It supports various outlier detection algorithms with unified interfaces. **TODS**[7] is a time-series outlier detection system with automated machine learning that provides pipeline search and hyperpa-

[5] github.com/taosdata/TDengine
[6] github.com/yzhao062/pyod
[7] github.com/datamllab/tods

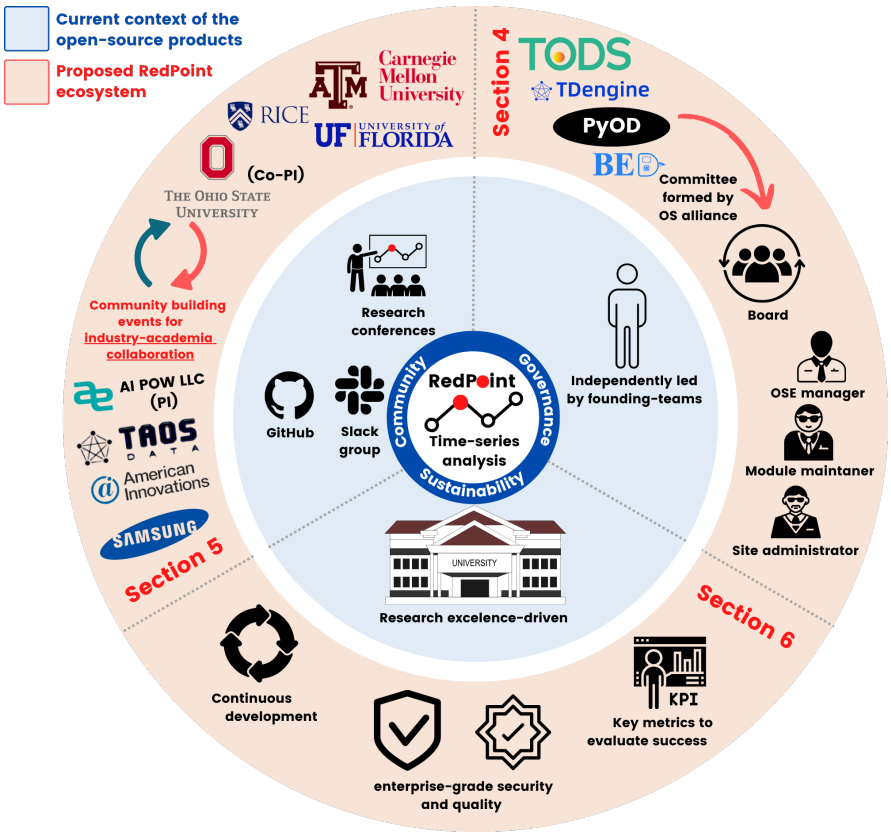

Figure 2: Context of the current open-source ecosystem focused on time-series analysis and sections of key activities proposed for Phase 2 RedPoint OSE for a sustainable open source ecosystem for **industry-academia collaboration** towards intelligent time series analysis.

rameter tuning to automatically identify the best model design. **B-E-D**[8], a system that provides solutions to the problem of deploying machine learning models in applications on edge devices like IoT sensors by using technologies for model compression.

In the same way, we have identified the following communities that are furthering the state-of-the-art in time-series analysis. **Data processing: Ts-fresh**[9] is another package that provides systematic time-series feature extraction functionaries. It includes statistics, signal processing, and nonlinear dynamics feature extraction modules. In addition, it provides feature section algorithms to forget irrelevant features. **SKTime**[10] is a Python library for time-series analysis. It supports a unified interface for multiple time-series processing. **Downstream applications: Ts-learn**[11] [48] is a general-purpose Python toolkit for time-series analysis. It provides various tools for pre-processing and feature extraction. Some basic machine learning models are also provided for standard tasks, such as classification and clustering. **sklearn**[12] is a general machine learning package, which also contains many functions for feature extraction. It also provides various machine learning modules for tasks such as classification and outlier detection.

[8] github.com/datamllab/BED_main
[9] github.com/blue-yonder/tsfresh
[10] github.com/sktime/sktime
[11] github.com/tslearn-team/tslearn
[12] github.com/scikit-learn/scikit-learn

Our goal is to develop a sustainable OSE based on a promising open-source product with the following innovation in the emergent machine learning-based time-series analysis. We can elaborate our progress from three aspects. **Ecosystem Development,** our team is continuously developing open-source libraries to attract users and contributors. **Research product,** we have published more than four papers in the related fields on top venues and corresponding toolkits. **Social Impact,** we have served as a contributor to the most impactful technical blog and host many discussions.

2.1.1 Current Status of Research

We have published several academic papers regarding time-series analysis, which will serve as the basis of our open-source community. Our research paper titled "Revisiting Time Series Outlier Detection: Definitions and Benchmarks" [51] has defined and studied different types of outliers in time-series data. Specifically, we proposed a behavior-driven taxonomy to clearly define the contexts of the outliers. Our taxonomy first categorizes time-series outliers as point-wise and pattern-wise outliers. The point-wise outlier is further categorized as global, contextual, and collective outliers, and the pattern-wise outliers are further divided into shapelet, seasonal, and trend outliers. Each type of outlier has clear context definitions. Based on the proposed behavior-driven taxonomy, we benchmark different types of algorithms. Our paper titled "Towards Similarity-Aware Time-Series Classification" [49] studied the time-series classification problem. This work examines the performance of traditional similarity-based methods and modern deep learning models for time-series classification under different amounts of supervision. Motivated by the various pros and cons of these different lines of research, we proposed a new framework that leverages graph neural networks to incorporate similarity information into the automated feature representation learning of deep neural networks. We demonstrated that our proposed algorithm outperformed other state-of-the-art time-series classification algorithms on the benchmark datasets. Our research and codes will serve as the code basis of the open-source ecosystem. Finally, the work on edge devices [50], is a system that reduces the size of machine learning models for deployment on edge devices like IoT sensors, and its currently developed to enabling time-series applications to run effitiently on IoT devices.

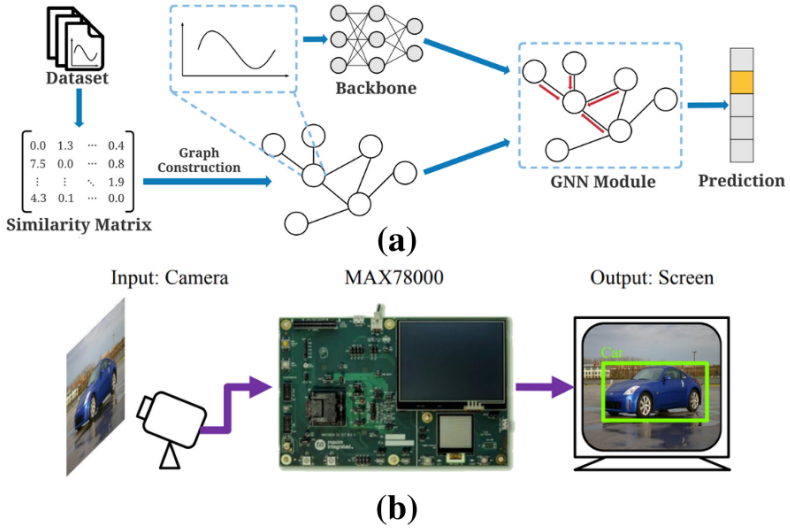

Figure 3: Examples of the current status of the research: (a) towards similarity-aware time-series classification [49] and (b) model compression research to fit time-series advanced analytics into edge devices [50].

2.1.2 Current Status of Open-Source Products

Several of our developed open-source projects have attracted the open-source community's attention. Their users will potentially become our first batch of contributors to our open-source ecosystem. First, an open-sourced system, named Automated Time Series Outlier Detection System (TODS) [52], is one of the most popular time-series outlier detection systems. It provides an exhaustive list of primitives for time-series processing, feature analysis, detection algorithms, and reinforcement, with more than 70 primitives. Users can flexibly create different types of pipelines based on the primitives. Our implementations of various preprocessing primitives will be our initial preprocessing modules in the ecosystem. We can redirect the users of TODS to our ecosystem. Our package AutoVideo [53] further extends the TODS to video analysis. Like time series, videos are also streaming data. Thus, many modules in AutoVideo could also be used to build the ecosystem. All our previous experiences and codes will be leveraged to build the proposed ecosystem. The users of our existing packages are our early users.

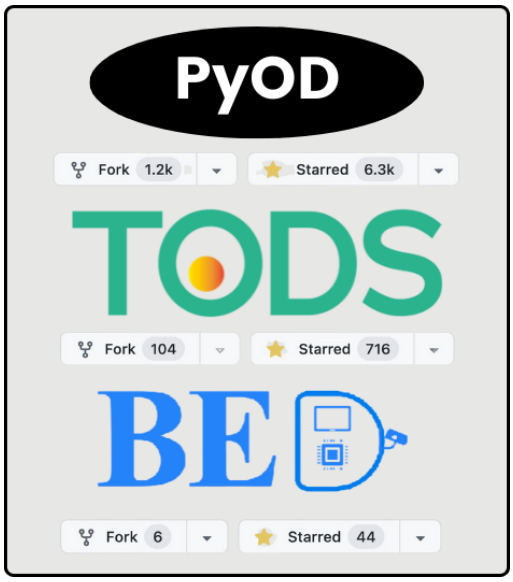

Figure 4: Examples of the current status of open-source products.

We have been closely working with our collaborators in CMU, and TDengine, who will help us promote the ecosystem. **Yue Zhao** from CMU is the author of PyOD [54], the most popular outlier detection system on Github with >6,300 stars, 1,200 forks, and 8,000,000 downloads. PyOD has a large user base, and outlier detection is a key application of PyOD. We will collaborate with the PyOD team to build our ecosystem. We will ensure that the feature extraction modules in our system are compatible with PyOD so that the users of PyOD can become part of the ecosystem.

2.2 Current Status of OSE Development

2.2.1 Social Impact

We have disseminated our OSE with three approaches: (1) open-source projects, (2) invited research talks, and (3) technical blog posts. Specifically, we have accumulated active users based on the developed open-source products (i.e., TODS, PyOD, and B-E-D). Furthermore, we have established multiple channels to allow broader discussions on the "Issues" in each project. Meanwhile, we have created corresponding slack and WeChat channels with more than 500 participants actively participating in related projects and can be deemed potential users of our OSE. In addition, we have been invited to academic conferences and industries to deliver research talks that are highly related to the OSE. We have delivered five talks at top-notch academic conferences and workshops (e.g., AAAI, NeurIPS, SDM) and three talks in tech companies (e.g., Samsung, Visa). The researchers and developers in those talks could be potential participants of our OSE and can be estimated to be more than 300 people. Furthermore, we actively contribute technical blog posts to TowardDataScience, one of the most popular data science technical publications with more than 600,000 followers with more than 14,000,000 monthly views. Our team has published four technical articles which attract 570,000+ readers and 100+ fans.

Figure 5: An illustration of the RAD development model.

2.2.2 User Base

The users of our OSE are mainly seeking to develop machine learning projects on their time series data, such as transaction logs, server logs, manufacturing line monitoring logs, or remote asset monitoring logs. The existing users are from both academia and industry. Specifically, researchers from universities and private research facilities seek to conduct new research and benchmarks based on our TODS, TDengine, B-E-D, and PyOD. **The projects from our OSE have obtained more than 6000 stars and 1200 forks on GitHub**. In addition, developers from different industries seek to develop products upon our OSE. To the best of our knowledge, our OSE has been applied to malware detection with companies such as Norton and Apple, manufacturing line monitoring with companies such as Ford Motors and General Motors, and remote asset monitoring with companies such as Trane.

2.3 Novelty of the Intended Product being Transitioned

The novelty of the proposed OSE lies in three aspects. First, our OSE focuses on vertical integration with various existing open-source projects. This way, domain experts can directly adopt the outcome of the OSE to real-world tasks in multiple industries. Second, our OSE develops a communication platform to connect academic researchers with industrial developers. Strong industry-academia collaboration can further accelerate industry development and provide opportunities for academic research on the most challenging real-world problems. Third, our OSE adopts a five-stage development procedure to ensure development quality and proposes various initiatives to drive sustainability in both academic research and engineering contributions to the OSE. More details will be elaborated on in the following sections.

3 Ecosystem Growth

Our team proposed three main tasks to grow the RedPoint ecosystem: software development, ecosystem dissemination, and partnership communication.

3.1 Software Development

To ensure the productivity of our software development, we plan to follow a five-stage procedure for development, management, and communication, illustrated in Figure 5. This procedure has the main focus on prototyping. Specifically, our team will allocate more effort to development instead of planning. This is very suitable for open-source packages because we need to quickly collect feedback from the users and the developers and iterate fast. In this way, we can deliver what we desire in a short period. The five stages are as follows: **1) User Modeling:** In this stage, we aim to understand users' needs for better management and planning. We will communicate with other open-source projects and do user surveys to understand what functionalities the users prefer and determine the priorities in the development. **2) Data Modeling:** This step aims to convert the information collected from the users into data objects, which involves joint development and management efforts. For example, we can map the user needs into classes, where each class is associated with some variables. These data objects will serve as the basis of our development.

3) Process Modeling: Given the data objects constructed in the data modeling stage, we aim to design the necessary workflow to achieve our needs, which is essential for project management. For example, we design how the data objects interact with each other with the class and instance methods. We define the input and output per functionalities and have a high-level design of how these functionalities are connected. **4) Application Development:** This stage converts the data objects and workflows into prototypes, which involves significant efforts of development. To save effort, we will use automation tools to generate the software and develop auto-generated software. **5) Prototype Testing:** After development, we test the prototype to ensure that it works as we desire. We will use automated testing scripts to generate a report in each iteration and reach out to potential users and people from other open-sourced projects to validate the design.

3.2 Ecosystem Dissemination

To disseminate the open-source community and related knowledge, we will host an organization on GitHub, the largest open-source community, conducting research, benchmark, and publish related blogs and research papers on impactful venues.

First, we propose establishing an organization to integrate existing open-source efforts on time series analysis and applications. Specifically, the organization will host various open-source projects ranging from time series data collection, pre-processing, and feature analysis to applications such as time series forecasting, anomaly detection, and classification. In addition, we will maintain a comprehensive list, named "awesome time series analysis," which will compose publicly available datasets, research papers, blogs, and code and projects to guide the beginner into this field. This way, the organization will become a friendly platform for practitioners to share and learn from each other.

Second, we propose continuously conducting research and benchmarks to truly push the community's progress. For example, we will collaborate with practitioners in the time series and related application communities to identify challenging problems for conducting research. The outcome of the research will be released, and the related development will be contributed to the proposed ecosystem. Furthermore, we will conduct comprehensive benchmarks on different applications, domains, and algorithms to make the community aware of the state-of-the-art and recent progress in academia. In the previous stage, we published two papers in top venues (i.e., AAAI and NeurIPS), which has led to numerous reads and more than 50 citations within one year. The outcomes of such research will attract people from academia and industry to deliver novel research and exciting problems to the field.

Third, we propose to continuously deliver technical blog posts to the developer communities. We will regularly write technical blogs to provide hands-on tutorials for the resources from our ecosystems. For instance, we will prepare step-by-step technical articles to trace the code of our projects, develop example applications based on the projects in our ecosystem, and explain complex ideas of the research fruits from our ecosystem to the general public. This way, researchers from academia will be willing to join the ecosystem to promote their work, and developers from the industry will have a chance to access to latest technologies.

3.3 Partnership Communication

To facilitate communications with potential collaborators, we will reach out to the related open-source project owners and contributors via email, WeChat groups, and a Slack channel called "Paper With Code." For example, the data mining and machine learning WeChat groups that are co-hosted with our collaborators Yue Zhao have more than 1500 people, and the Slack channel has more than 17,000 open-source contributors.

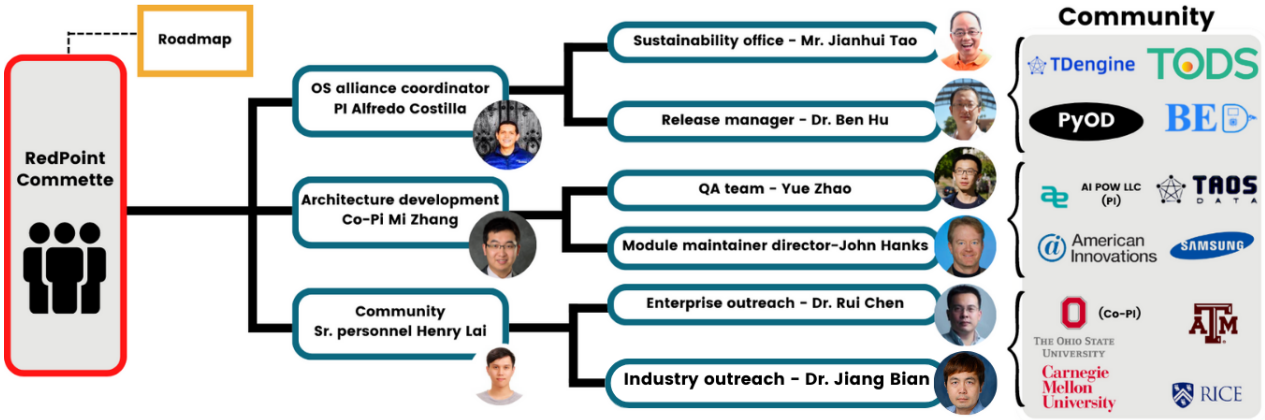

Figure 6: RedPoint organization structure as an evolution of the previous founder-team-led structure.

In addition, to encourage the participation of existing open-source projects in our OSE, we will provide concrete benefits and initiatives. First, the project owner will be able to keep their intellectual property. We will establish a GitHub organization to integrate the participant projects and assign a badge of our OSE to the participant projects. The assigned badge will help us to create a portal to direct the traffic of related projects in our OSE to the participant projects. Second, we will provide resources for project management and continuous development and maintenance (details are described in Section 4.2 and Section 3.1). Third, our OSE will provide various opportunities to broaden the audience of participant projects (details described in Section 5).

4 Organization and Governance

In our current founder-team-led governance model, our group who started the project also administers the project, establishes its vision, and controls permissions to merge code into it. Under this model, the PyOD, TDengine, and TODS founders are the final decision-makers for project matters. Unfortunately, the founder-team-led model has become a bottleneck for project decision-making work. We have been experiencing this with the TODS project, and e want to ensure that we can continue growing and scaling as a team. To do this, we need a governance structure that allows us to make decisions as a group while still empowering each individual in our group to be fully engaged with the project. We propose a committee-led governance structure that will allow us to create an open decision-making process among all contributors. Under this model, we are creating a new ecosystem structure *to manage the ecosystem (PI Alfredo Costilla-Reyes), OSE architecture development and deployment (Co-PI Mi Zhang), community building activities (Senior Personnel Henry Lai), and a general OSE committee.*

Figure 6 shows the different roles and members of the organization structure planned for the OSE RedPoing along with their specific community building activities (described in Section 5).

PI Alfredo Costilla-Reyes will be the OS alliance coordinator at RedPoint's. He is serving as Chief Executive Officer at AI POW LLC, a Hispanic-led company and spin-off from Rice University (Department of Computer Science). Dr. Costilla is leading the product R&D in time-series anomaly detection and explainable AI for IoT devices. He is also the **first-generation graduate** from both the Entrepreneurship and Technology Commercialization program at Mays Business School and the doctorate program in Electrical Engineering, both from Texas A&M University.

Co-PI Mi Zhang will serve as RedPoint's architecture development and deployment. He is also an Associate Professor at The Ohio State University. He is an expert in AutoML, model compression, and AIoT. He is the 4th Place Winner (1st Place in U.S. and Canada) of the 2019

Google MicroNet Challenge. The neural network model compression and AutoML techniques his team developed are the key enablers for running deep learning-based applications on resource-constrained IoT devices. He is the Third Place Winner of the 2017 NSF Hearables Challenge and the 2016 NIH Pill Image Recognition Challenge champion. In addition, his works won seven best paper awards and nominations. He is also the recipient of the NSF CRII Award, Facebook Faculty Research Award, and Amazon Machine Learning Research Award.

Senior Personnel Henry Lai will take the community-building responsibilities at RedPoint. Henry is currently the Vice President of AI POW's open-source initiative. Henry has previously worked at Visa and Academia Sinica and established joint research with KKBOX, KKTIX, Cathay United Bank, General Motors, and Trane, providing cutting-edge and large-scale machine-learning solutions for the industry. His works have directly impacted production systems, including multimedia recommender systems, financial fraud detectors, manufacturing lines, and remote asset monitoring systems. Those impacts have also led to 6 U.S. patents and various publications on top venues. He has developed four popular machine learning-related open-source projects, which obtained more than 3200+ stars and 600+ forks.

Senior Personnel Dr. Xia "Ben" Hu will be RedPoints new version release manager. He is also an Associate Professor at Rice University in the Department of Computer Science. He has published more than 100 papers in major data mining venues. His articles have received seven Best Paper Awards (candidate), and he is the recipient of the JP Morgan AI Faculty Award, the Adobe Data Science Award, and the NSF CAREER Award. His group's open-source package, AutoKeras, has become the most used automated deep learning system on Github (with over 8,500 stars and 1,400 forks). Dr. Hu's work on deep collaborative filtering, anomaly detection, and knowledge graphs is part of the TensorFlow package, Apple production system, and Bing production system.

Collaborator Yue Zhao will be in charge of the quality assurance team. He has made enormous contributions to reproducible, automated, and scalable outlier detection: He uses machine learning systems (MLSys) techniques to support large-scale, real-world outlier detection applications. His work has been widely used by thousands of projects and applications, including Amazon, IBM, Morgan Stanley, and Tesla.

Collaborator Jianhui Tao will use his background to lead the sustainability office. He is a Serial entrepreneur and founder of TDengine, iCareNewLife, and Hesine Inc. He is a strong believer in the power of technology and has dedicated himself to applying software intelligence to improve our lives. Mr. Jeff is passionate about open source, technology, and innovation—and his primary contribution to the field has been as a core developer of TDengine, an open-source cloud-native time series database.

Collaborator Dr. Rui Chen he is the Chief Scientist, Senior Director, and Senior Principal Research Scientist, at Samsung Research America. His focuses on production-scale deep/machine learning platforms and algorithms for real-world applications, including digital advertising, personalized recommendation, natural language understanding, and user profiling. Dr. Chen will help the RedPoint acosystem with the enterprise outreach.

Collaborator Dr. Jiang Bian will be our industry outreach officer. He is a Professor and Chief Data Scientist at University of Florida Health. He will bring his expertise in the healthcare domain to work with the community to create an open-source solution that helps healthcare professionals access intelligent time-series analysis and innovate with their data assets by providing curated data, real case studies, and feedback.

Collaborator John Hanks will coordinate the module maintainer efforts. He is an experienced industry veteran with more than 30 years of corporate experience at NI, Maxim Integrated and Siemens Medical Systems with roles leading engineering, R&D and product management teams.

4.1 Intellectual Property

Our team will release the intellectual property generated in this project under the open software license [55]. This license type will allow us to open-source our core system with basic functionalities to enable potential customers to build their anomaly detection system. Specifically, the system core functions derived from D3M [56] and essential functions, which allow users to create a basic anomaly detection system, will be published with Apache License 2.0. This license allows users to use RedPoint for any purpose without restrictions. However, this license does not allow users to modify RedPoint and develop a competing product. The project team will provide documentation on how to implement the core system functions and use the data provided by our team to train models. In addition, we will publish the research results of this project under the Creative Commons Attribution License 4.0. This license type allows users to use our research results for any purpose as long as they give credit to the original authors. The goal is to encourage academic researchers to conduct research based on RedPoint.

4.2 Continuous Development

Our continuous development strategy involves keeping employing GitHub as our primary environment, where new features can be added to RedPoint as soon as they are ready. We will implement the following measures to ensure the sustainability of RedPoint: **1) CI/CD:** A continuous deployment pipeline with automated tests involves Continuous integration (CI) and Continuous Delivery (CD). These tools allow us to continuously deploy new versions of the application to production, so we can quickly respond to user feedback. With CD, every change made in development is automatically tested on a staging environment where any problems will be identified before they are deployed into production. **2) Documentation:** Documentation for each new release will allow our users to understand how to use the new features and will make it easier for us to manage the application. We will also have documentation of our code base that can be used to help new developers who join the team. **3) Bug-free prevention:** The bug tracking and reporting system are intended to be used by our development team to keep track of the bugs found in each release and for our users to report any problems they encounter. This will help us identify issues quickly to fix them before they affect too many users. **4) Quality control:** A QA team will ensure the quality of key features of RedPoint by testing new releases to ensure they work as expected and identifying any bugs the developer team may have missed. They will also test new features before they are released to ensure they do not cause any problems with existing functionality.

The measures mentioned above are designed for a highly functioning asynchronous and distributed development of RedPoint. RedPoint is an ambitious ecosystem, and we must ensure the quality of each release. As RedPoint grows and more developers join us, the development team's size will also increase. We want to make sure that everyone on our team knows what their responsibilities are when it comes to testing new releases and reporting any issues they may find. This will help us identify problems early when they are easier and cheaper to fix.

Our metrics to assess and evaluate success will be based on the quality of our releases, measured by the number of bugs reported in each release and the number of new developers who join RedPoint and start contributing code and documentation. We further elaborate on sustainability metrics for RedPoint in Section 6.

Finally, to secure the privacy of new content, we have established a security team responsible for ensuring that the data we collect is secure and that it's not shared with anyone outside of RedPoint. This includes ensuring our servers are secure and that all changes to our codebase follow the appropriate procedures.

5 Community Building

Our current community-building activities that have been very successful include blogs and a Slack channel that we use for communicating with our community members. We also have a team of volunteers working on tutorials, documentation, and training materials to help new users get up to speed more quickly. However, for this new phase of ecosystem growth, we want to expand our efforts and have a more formalized program for community building for our users to interact with the platform and its community. Our goals are to make it easier for users to find and connect with each other and provide more opportunities for them to contribute back to the ecosystem by submitting bug reports or helping new users get started.

The specific activities will aid in developing a strategy to engage potential content contributors who will help build and maintain the open-source product. Figure 7 shows a metric for the level of user involvement for the planned community building (CB) mechanisms to actively engage with our communities below.

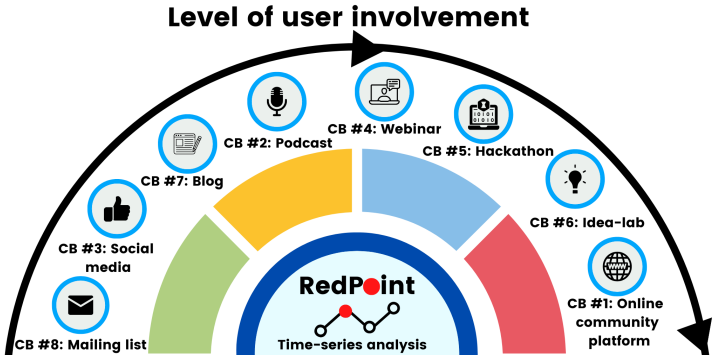

Figure 7: Planned activities for community building activities and level of user involvement metric.

Launch of an **CB #1 online community platform** for our contributors to share ideas, provide feedback, and collaborate with code and content. This platform will centralize the different time-series analysis tools available on GitHub and provide the different planned events and channels users, and developers can interact at RedPoint i.e., Slack channel and WeChat group. This will include demo applications, documentation for getting started, and a FAQ section for common questions. **We plan to launch this platform by June 2023.**

CB #2 Podcast: We will launch a podcast series to educate our community on time-series analysis and data science. The podcast will be hosted by RedPoint's community managers and engineers, who can provide insight into their tools' design process and share best practices for using them. We plan to publish the first episode within three months of launching our community program. This podcast will be available on iTunes and Google Play Music for easy access by anyone interested in learning about time-series analysis. **We expect to produce the first podcast episode by August 2023 and publish one episode per month after that.**

CB #3 Social media: We will launch an official RedPoint social media campaign to engage our audience and promote our ecosystem. Social media is an effective way to reach out to users, so we want to make it easy for people interested in learning more about time-series analysis and data science to find us. Our original campaign will only include LinkedIn and Twitter posts for now, and **it will be launched by June 2023.**

CB #4 Webinars: We will provide valuable industry insights to help with workshops programmed for the general public. The sessions we have planned will cover introductory topics on time-series analysis for new users, as well as a more specific program that covers particular use cases in industries such as manufactury, energy, industrial Internet of Things (IoT), and remote asset monitoring for medium and advanced users that require time-series analysis in their day-to-day data science work. **We plan to produce five webinars throughout 2023 and 2024.**

CB #5 Hackathon: We will host a hackathon that will be fully focused on the proposed OSE-TODS. A competition organized in three major US Universities will allow the proposing team to

demonstrate the interest of the academic and industrial community in the OSE. The hackathon will be a great opportunity to evaluate the ease of access and use of the proposed OSE, RedPoint. It will also allow us to better understand the needs of our users and improve the software. **Our first and second hackathon will be scheduled for January 2024 and January 2025.** The tentative places to host the Hackathon and their corresponding topic are as follows.

- □ The Ohio State University. Time-series on edge devices (resource constrained hardware) for defect detection in the semiconductor industry.
- □ Carnegie Mellon University. RedPoint and cybersecurity.
- □ Rice University. Time-series analysis for health applications.
- □ Texas A&M University. Using RedPoint to solve challenges in the oil and gas industry.

CB #6 Idea Labs: We intend to host three Idea Lab that intends to attract domain experts in the following industries: energy, health, and manufacturing. This 'RedPoint Idea Lab' is programmed to be a five-day intensive and interactive work with diverse participants from the energy, health, and manufacturing industries to immerse in collaborative thinking processes to construct innovative approaches to use time-series analysis tools in their domains. We expect to invite 35 participants to identify and define the scope of the challenges relating to forming and sustaining an OSE based on RedPoint suite of open-source projects. **We have scheduled the topic-focused idea labs as follows: energy on February 2023, health on June 2023, and manufacturing on October 2023. Topics for the idea labs of 2024 will be selected at the end of each idea lab of 2023.** The tentative places to host the idea lab and their corresponding topic are as follows.

- □ American Innovations in Houston, Texas, with the topic "Pipeline infrastructure safety and compliance through time-series data on edge devices."
- □ TAOS data in Sacramento, California, with the topic "Time-series databases for the future internet of things."
- □ Samsung at San Francisco, California, "time-series data in remote asset monitoring".

CB #7 Blog: We plan to launch a blog that will publish industry-wide news, important developments in the time-series analysis space, and educational content. This blog will be hosted on our RedPoint ecosystem website and Medium. **We plan to publish our first RedPoint blog by October 2023 and then bi-monthly afterward.**

CB #8 Mailing list: We have created a mailing list to facilitate communication between the developers and users to discuss the project's development and help new contributors get started. Our first task on the mailing list is to announce the ecosystem, its goals, and the community-building activities planned for the future. We will also use this mailing list as a preferred method for future communications about the project's development. We want to stress that we have already created a **Slack community** in addition to our mailing list, which allows users to connect and exchange ideas. We will use every planned community activity to onboard potential content contributors by providing them with the opportunity to contribute to the ecosystem and by providing them with a clear path for how they can get started. This will be a long-term effort, but we believe it is an important step in building our community. We want to incentivize new members that join RedPoint and avoid one of the most common pitfalls of open-source projects: content contributors who only contribute once or twice and disappear. For this, we plan to launch a "contribution bounty system" for our community members tailored to each of the community-building activities mentioned above. This will be a way to incentivize newcomers who want to contribute content but don't know where or how to start. We will provide them with a list of currently missing tasks from the ecosystem and explain how they can get started with these tasks. In

this new community-driven content contribution model, we will reward contributors with badges and points for completing tasks on the list. The badge system will allow us to recognize and reward contributors with different contribution levels. It will also help us identify the most active community members and recognize them for their efforts.

6 Sustainable Business Model

The open-source model is the best path to software development because it speeds up product feedback and innovation, improves software reliability, scales support, drives adoption, and pools technical talent. The open-source model has proven successful in almost every industry, and it's no surprise that the software world is now embracing it. However, the full potential of open-source is only realized when technological innovation is paired with commercial innovation. While open source has a great track record of creating large-scale software, these projects often lack the tools to create a sustainable business model. The result is that many of these projects don't survive long enough to grow their community or reach their full potential for success.

HuggingFace Inc serves as our model of a community-driven repository for open-source ML technology. The company was founded in 2016 to improve the quality and availability of other people's work by providing it in a well-documented, easy-to-use format. HuggingFace built an online platform allowing users to create their portfolios of code snippets that are then shared with others. Users can submit their code snippets so others can use them. HuggingFace allows users to contribute financially towards projects they like through donations on their platform; however, this isn't required since most content is freely available without any monetary obligations.

Following some guidelines from successful open-source communities such as HuggingFace, we understand for RedPoint to thrive long-term, we must ensure it is stable and reliable enough to meet their needs. Therefore, to preserve the quality of the open-source ecosystem, we follow the common practice of CI/CD (Continuous Integration / Continuous Deployment). We will develop automation scripts to enable continuous integration. First, we support integrated/unit tests to cover most functionalities. Once a new code is merged into the codebase, the testing scripts will be automatically triggered to ensure that added new codes will not cause issues. Second, we include a building test, which will also be automatically triggered to ensure that the new code can be built and run in different programming environments. The building test script will cover different versions of the programming languages and libraries. To easily process the issues and pull requests, we will use auto-bot to do auto filtering, which could help us identify the most important issues. To support continuous deployment, we follow standard release and deployment procedures. We have a regular release cycle of the codebase to release on a weekly or monthly basis. The maintainers, developers, and users can easily track different package versions and roll back if something gets wrong. Putting all the above together, we aim to ensure the quality of the open-sourced ecosystem, enabling reproducible and scalable model training.

Lastly, we believe that the best way to build an effective open-source project is by having community involvement and commercial backing. Therefore, we want to pay special attention to the need for commercial support. Without it, there will likely be challenges in funding RedPoint and maintaining its sustainability after this proposed POSE phase 2 work. Therefore, we will explore an open-source business model by offering paid training. In parallel with CB #4, we plan to offer industry-specific webinars that may interest the attendees of our planned community-building plans. It is important to stress that the open-source code in RedPoint will always be free and only specific features requested training will be paid. We want to ensure that we are not creating a barrier for the community to use RedPoint but rather offering an opportunity for those interested in learning more about commercial support for open-source software.

7 Metrics to Evaluate Success

The metrics we will implement to assess and evaluate success are as follows:

- □ **Number of users who use our tools and libraries for time-series analysis for personal, academic, and industrial uses.** This metric will be tracked through our proposed online community platform (CB #1). Users will be able to earn 'badges' to encourage them to contribute new code and help maintain the open-source repository by fixing bugs.

- □ **Number of developers and maintainers contributing to our open-source projects and how many of them are new contributors.** We intend that the most active contributors are also invited to contribute to the community-building activities CB #4 and #7. New contributors will be asked how did they hear about RedPoint to measure the effectiveness of community-building activities CB #2, #3, and #7.

- □ **Number of users who have issues with our tools and libraries and how many of them are successfully resolved within two weeks (if possible).** This metric will be tracked primarily via GitHub's available developer metrics.

- □ **Number of users who employ our tools and libraries on GitHub for time-series analysis in their research papers, publications, or industrial applications.** While Google scholar is the best method to track the academic impact of RedPoint, we will use community-building activities CB #5 and #6 to understand how RedPoint is used in industry and to measure its impact beyond academia.

8 Broader Impacts

This POSE project will enhance partnerships **between academia and small businesses with strong research capabilities in the US.** We firmly believe that by partnering with the Hispanic-led startup AI POW LLC, we will be driving diversity, excellence, and innovation among our research members and third-party collaborations. The POSE project will bring together a diverse group of researchers from industry and academia to create a community of practice around time series data and AI in the cloud. The project will also provide opportunities for industry-driven research that has the potential to benefit both academia and industry partners.

Harnessing the Data Revolution. As data science and engineering become more critical for businesses, and it is paramount to consolidate a 21st-century data-capable workforce. Currently, many large enterprises can implement analysis detection systems because they can afford a highly specialized computer scientist team. This project aims to democratize easy-to-use tools tailored to software engineers with basic ML knowledge, **and accelerate the technology adoption for a large group of small and medium-sized businesses that can now embrace the data revolution** to start using their data assets to innovate in new ways.

Future of Work at the Human-Technology Frontier. More companies and researchers could use complex anomaly detection frameworks by increasing the adoption of the proposed advanced AI technologies. Successful human-technology partnerships could effectively enhance human performance to simplify excessive amounts of data into actionable items to boost industrial equipment efficiency through predictive maintenance procedures determining when a remote asset is about to fail or detect a defective part. Prompt and systematic error detection represents actual money and time savings in equipment repairs to a company and ensures standardization in production, leading to higher product quality.

References

[1] Markets and Markets. *US Industry Report Code: TC 5109. Remote Asset Management Market*, May 2020. Retrieved from Markets and Markets database.

[2] Dan Cook. *IBISWorld US Industry (NAICS) Report 51121C. Business Analytics Enterprise Software Publishing in the US*, July 2020. Retrieved from IBISWorld database.

[3] Spyros Makridakis. A survey of time series. *International Statistical Review/Revue Internationale de Statistique*, pages 29–70, 1976.

[4] Martin Längkvist, Lars Karlsson, and Amy Loutfi. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42:11–24, 2014.

[5] Ruey S Tsay. *Analysis of financial time series*. John wiley & sons, 2005.

[6] Terence C Mills and Raphael N Markellos. *The econometric modelling of financial time series*. Cambridge university press, 2008.

[7] Sabyasachi Basu and Martin Meckesheimer. Automatic outlier detection for time series: an application to sensor data. *Knowledge and Information Systems*, 11(2):137–154, 2007.

[8] Siddhartha Bhandari, Neil Bergmann, Raja Jurdak, and Branislav Kusy. Time series data analysis of wireless sensor network measurements of temperature. *Sensors*, 17(6):1221, 2017.

[9] Karsten Sternickel. Automatic pattern recognition in ecg time series. *Computer methods and programs in biomedicine*, 68(2):109–115, 2002.

[10] Klaus Lehnertz. Non-linear time series analysis of intracranial eeg recordings in patients with epilepsy—an overview. *International Journal of Psychophysiology*, 34(1):45–52, 1999.

[11] James Douglas Hamilton. *Time series analysis*. Princeton university press, 2020.

[12] Chris Chatfield. *Time-series forecasting*. Chapman and Hall/CRC, 2000.

[13] Ganapathy Mahalakshmi, S Sridevi, and Shyamsundar Rajaram. A survey on forecasting of time series data. In *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, pages 1–8. IEEE, 2016.

[14] Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.

[15] Ali Azadeh, SF Ghaderi, and S Sohrabkhani. Forecasting electrical consumption by integration of neural network, time series and anova. *Applied Mathematics and Computation*, 186(2):1753–1761, 2007.

[16] David L McCollum. Machine learning for energy projections. *Nature Energy*, 6(2):121–122, 2021.

[17] Gul Muhammad Khan, Jawad Ali, and Sahibzada Ali Mahmud. Wind power forecasting—an application of machine learning in renewable energy. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 1130–1137. IEEE, 2014.

[18] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963, 2019.

[19] Pierre Geurts. Pattern extraction for time series classification. In *European conference on principles of data mining and knowledge discovery*, pages 115–127. Springer, 2001.

[20] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery*, 31(3):606–660, 2017.

[21] Argyro Kampouraki, George Manis, and Christophoros Nikou. Heartbeat time series classification with support vector machines. *IEEE transactions on information technology in biomedicine*, 13(4):512–518, 2008.

[22] Wanpracha Art Chaovalitwongse, Oleg A Prokopyev, and Panos M Pardalos. Electroencephalogram (eeg) time series classification: Applications in epilepsy. *Annals of Operations Research*, 148(1):227–250, 2006.

[23] Ting Wang, Sheng-Uei Guan, Ka Lok Man, and TO Ting. Eeg eye state identification using incremental attribute learning with time-series classification. *Mathematical Problems in Engineering*, 2014, 2014.

[24] Bovas Abraham and Alice Chuang. Outlier detection and time series modeling. *Technometrics*, 31(2):241–248, 1989.

[25] Zakia Ferdousi and Akira Maeda. Unsupervised outlier detection in time series data. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages x121–x121. IEEE, 2006.

[26] Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. Fraud detection system: A survey. *Journal of Network and Computer Applications*, 68:90–113, 2016.

[27] Yufeng Kou, Chang-Tien Lu, Sirirat Sirwongwattana, and Yo-Ping Huang. Survey of fraud detection techniques. In *IEEE International Conference on Networking, Sensing and Control, 2004*, volume 2, pages 749–754. IEEE, 2004.

[28] Dubravko Miljković. Fault detection methods: A literature survey. In *2011 Proceedings of the 34th international convention MIPRO*, pages 750–755. IEEE, 2011.

[29] Jenny A Harding, Muhammad Shahbaz, and A Kusiak. Data mining in manufacturing: a review. 2006.

[30] Alok Kumar Choudhary, Jenny A Harding, and Manoj Kumar Tiwari. Data mining in manufacturing: a review based on the kind of knowledge. *Journal of Intelligent Manufacturing*, 20(5):501–521, 2009.

[31] Karanjit Singh and Shuchita Upadhyaya. Outlier detection: applications and techniques. *International Journal of Computer Science Issues (IJCSI)*, 9(1):307, 2012.

[32] Ashraf Darwish and Aboul Ella Hassanien. Wearable and implantable wireless sensor network solutions for healthcare monitoring. *Sensors*, 11(6):5561–5595, 2011.

[33] Sandeep Kumar Vashist, E Marion Schneider, and John HT Luong. Commercial smartphone-based devices and smart applications for personalized healthcare monitoring and management. *Diagnostics*, 4(3):104–128, 2014.

[34] Stefano Longo, Benedetto Mirko d'Antoni, Michael Bongards, Antonio Chaparro, Andreas Cronrath, Francesco Fatone, Juan M Lema, Miguel Mauricio-Iglesias, Ana Soares, and Almudena Hospido. Monitoring and diagnosis of energy consumption in wastewater treatment plants. a state of the art and proposals for improvement. *Applied energy*, 179:1251–1268, 2016.

[35] Hardik A Gohel, Himanshu Upadhyay, Leonel Lagos, Kevin Cooper, and Andrew Sanzetenea. Predictive maintenance architecture development for nuclear infrastructure using machine learning. *Nuclear Engineering and Technology*, 52(7):1436–1442, 2020.

[36] Seyed Mahdi Miraftabzadeh, Federica Foiadelli, Michela Longo, and Marco Pasetti. A survey of machine learning applications for power system analytics. In *2019 IEEE International Conference on Environment and Electrical Engineering and 2019 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe)*, pages 1–5. IEEE, 2019.

[37] Martin G Schultz, Clara Betancourt, Bing Gong, Felix Kleinert, Michael Langguth, Lukas Hubert Leufen, Amirpasha Mozaffari, and Scarlet Stadtler. Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A*, 379(2194):20200097, 2021.

[38] Christy Kunjumon, Sreelekshmi S Nair, Padma Suresh, SL Preetha, et al. Survey on weather forecasting using data mining. In *2018 Conference on Emerging Devices and Smart Systems (ICEDSS)*, pages 262–264. IEEE, 2018.

[39] Agustín Agüera-Pérez, José Carlos Palomares-Salas, Juan José González De La Rosa, and Olivia Florencias-Oliveros. Weather forecasts for microgrid energy management: Review, discussion and recommendations. *Applied energy*, 228:265–278, 2018.

[40] Ziqiu Kang, Cagatay Catal, and Bedir Tekinerdogan. Machine learning applications in production lines: A systematic literature review. *Computers & Industrial Engineering*, 149:106773, 2020.

[41] Dorina Weichert, Patrick Link, Anke Stoll, Stefan Rüping, Steffen Ihlenfeldt, and Stefan Wrobel. A review of machine learning for the optimization of production processes. *The International Journal of Advanced Manufacturing Technology*, 104(5):1889–1902, 2019.

[42] Michael Sharp, Ronay Ak, and Thomas Hedberg Jr. A survey of the advancing use and development of machine learning in smart manufacturing. *Journal of manufacturing systems*, 48:170–179, 2018.

[43] Xun Zhou, Sicong Cheng, Meng Zhu, Chengkun Guo, Sida Zhou, Peng Xu, Zhenghua Xue, and Weishi Zhang. A state of the art survey of data mining-based fraud detection and credit scoring. In *MATEC Web of Conferences*, volume 189, page 03002. EDP Sciences, 2018.

[44] Khaled Gubran Al-Hashedi and Pritheega Magalingam. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, 40:100402, 2021.

[45] Atif Alamri, Wasai Shadab Ansari, Mohammad Mehedi Hassan, M Shamim Hossain, Abdulhameed Alelaiwi, and M Anwar Hossain. A survey on sensor-cloud: architecture, applications, and approaches. *International Journal of Distributed Sensor Networks*, 9(2):917923, 2013.

[46] Mohammed Talal, AA Zaidan, BB Zaidan, Ahmed Shihab Albahri, Abdullah Hussein Alamoodi, Osamah Shihab Albahri, MA Alsalem, Chen Kim Lim, Kian Lam Tan, WL Shir, et al. Smart home-based iot for real-time and secure remote health monitoring of triage and priority system using body sensors: Multi-driven systematic review. *Journal of medical systems*, 43(3):1–34, 2019.

[47] Mohammad Saeid Mahdavinejad, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi, and Amit P Sheth. Machine learning for internet of things data analysis: A survey. *Digital Communications and Networks*, 4(3):161–175, 2018.

[48] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, et al. Tslearn, a machine learning toolkit for time series data. *J. Mach. Learn. Res.*, 21(118):1–6, 2020.

[49] Daochen Zha, Kwei-Herng Lai, Kaixiong Zhou, and Xia Hu. Towards similarity-aware time-series classification. In *Proceedings of the 2017 SIAM international conference on data mining*, 2022.

[50] Guanchu Wang, Zaid Pervaiz Bhat, Zhimeng Jiang, Yi-Wei Chen, Daochen Zha, Alfredo Costilla Reyes, Afshin Niktash, Gorkem Ulkar, Erman Okman, and Xia Hu. Bed: A real-time object detection system for edge devices, 2022.

[51] Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Hu. Revisiting time series outlier detection: Definitions and benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[52] Kwei-Herng Lai, Daochen Zha, Guanchu Wang, Junjie Xu, Yue Zhao, Devesh Kumar, Yile Chen, Purav Zumkhawaka, Minyang Wan, Diego Martinez, et al. Tods: An automated time series outlier detection system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 16060–16062, 2021.

[53] Daochen Zha, Zaid Pervaiz Bhat, Yi-Wei Chen, Yicheng Wang, Sirui Ding, Jiaben Chen, Kwei-Herng Lai, Anmoll Kumar Jain, Mohammad Qazim Bhat, Na Zou, et al. Autovideo: An automated video action recognition system. *arXiv preprint arXiv:2108.04212*, 2021.

[54] Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *arXiv preprint arXiv:1901.01588*, 2019.

[55] Lawrence Rosen. *Open source licensing*, volume 692. 2005.

[56] Diego Martinez-Garcia Saswati Ray Sujen Shah Mitar Milutinovic, Brandon Schoenfeld and David Yan. On evaluation of automl systems. In *ICML Workshop on Automated Machine Learning*, 2020.