Securities & Exchange Organization, Research, Development & Islamic Studies (RDIS)
Journal of Securities and Exchange, Winter 2025, V. 18, No.72, pp. 241-268

A Method for Increasing the Accuracy of Credit Imbalanced Data[1]

Arash GhorbanniaDelavar[2], Sadaf Sadat Ziya[3]

Receive Date:29 October 2022 Revise Date: 21 December 2025
Accept Date: 21 December 2025 Publish Date: 21 December 2025

Research Paper

Highlights

- This study introduces a hybrid computational approach that combines Support Vector Machine (SVM) and Deep Belief Network (DBN) techniques for the classification of imbalanced credit data.
- It presents a two-phase optimization pipeline, which encompasses "feature selection" and "kernel Modification" aimed at minimizing data redundancy and duplication.
- The focus is placed on the "Weighted Accuracy" (Waccuracy) metric to effectively tackle class imbalance and the uneven costs associated with credit risk evaluation.
- The proposed methodology resulted in an increase of approximately 7.5% in mean weighted accuracy for the German dataset and around 2% for the Japanese dataset.
- The research illustrates that the optimization pipeline (FS + Kernel) serves as a robust, model-agnostic framework, enhancing the performance of various classification models.

Abstract

The main goal of this research is to provide a method that can be used to increase the accuracy of credit imbalance data. Financial fraud is a fundamental problem that affects both the financial sector and life and plays an important role in affecting the integrity and trust in the financial sector, as well as the cost of living. Data classification in support vector machine plays a very important role in increasing accuracy. For this purpose, we will present a method for credit imbalance data classification using support vector machine and deep belief network. With a case study on the methods of deep belief network and support vector machine, unbalanced credit data has not been taken into consideration, and there has also been data redundancy and duplicate data in these methods. We want to reduce data redundancy and duplicate data and finally increase the weight of data accuracy for credit imbalanced data using feature selection and kernel modification method.

Keywords: Imbalanced Data, Credit Imbalanced Data, Credit Classification, Accuracy of Imbalanced Data, Accuracy.
JEL Classification: E5, E51, C10, C55.

---

1. doi: 10.22034/JSE.2025.12096.2066

2. Associate Professor, Department of Computer Engineering, Faculty of Engineering, Payame Noor University, Tehran, Iran. (Corresponding Author). (a_ghorbannia@pnu.ac.ir).
3. M.Sc. Graduate, Department of Computer Engineering, Faculty of Engineering, Payame Noor University, Tehran, Iran. (zia.s69@gmail.com).

Cite This Article: GhorbanniaDelavar, Arash; Sadat Ziya, Sadaf. (2025). A Method for Increasing the Accuracy of Credit Imbalanced Data. *Journal of Securities Exchange*, 18 (72), 241-268. http://10.22034/JSE.2025.12096.2066.

Introduction

The research paper titled "A Method for Increasing the Accuracy of Credit Imbalanced Data" introduces a hybrid computational technique aimed at addressing the significant issue of classification accuracy within imbalanced credit datasets. Financial fraud and credit risk are critical challenges faced by the global financial industry. The authors highlight a significant research gap: although models such as Support Vector Machines (SVM) and Deep Belief Networks (DBN) have been utilized in credit scoring, previous implementations have frequently fallen short in effectively addressing the specific issue of class imbalance. Additionally, these applications have been hindered by data redundancy and duplicate entries, which adversely affect model performance. The main aim of this research is to propose an innovative classification method that merges SVM and DBN. This hybrid strategy is further refined through two essential optimization processes: a systematic feature selection approach and a kernel modification technique, specifically designed to optimize the Radial Basis Function (RBF) kernel. The ultimate objective is to diminish data redundancy and enhance the weighted accuracy of classifications for imbalanced credit data.

Methodology

This paper positions its research within the framework of a fluctuating contemporary economy where effective credit risk assessment is crucial. Conventional risk models are criticized for being "one-dimensional" frequently depending on simplistic data and lacking the ability for thorough, multi-faceted analysis, which leads to diminished accuracy. Data mining methodologies are suggested as a more holistic alternative.

A significant obstacle in this field is the uneven cost associated with misclassification. The expense of a False Negative (FN)—approving a detrimental customer and forfeiting the entire principal—is alarmingly greater than the cost of a False Positive (FP)—denying a beneficial customer and losing merely potential interest. This economic disparity makes basic accuracy an inadequate measure. The model must be refined to emphasize the reduction of FNs, which is the reason the paper concentrates on metrics that are attuned to this imbalance, such as Waccuracy (Weighted Accuracy), which offers a more equitable

evaluation of a model's efficacy concerning the critically significant minority class (undesirable customers).

The study builds on a well-established corpus of research in machine learning pertaining to imbalanced data. "Theoretical Foundations" segment of the paper classifies existing solutions into four main categories:

1. Data-Level Methods: Modifying the data distribution (e.g; over-sampling or under-sampling).
2. Algorithm-Level Methods: Adjusting the classification algorithms themselves to enhance sensitivity to the minority class.
3. Cost-Based Methods: Imposing a greater penalty for the misclassification of the minority class.
4. Hybrid Approaches: Merging several weak classifiers to form a single, robust model.

The paper outlines the progression of models in credit scoring, transitioning from traditional algorithms such as Decision Trees (DT) and SVM, to more contemporary, sophisticated techniques like Gradient Boosted Decision Trees (GBDT).

The authors propose a hybrid model, referred to as IDCOST in the results, which is based on a combination of Support Vector Machine (SVM) and Deep Belief Network (DBN) and incorporates a bagging algorithm.

However, the paper's core innovation is not a novel model architecture. Instead, the authors identify the primary failure of previous SVM and DBN applications as data redundancy and duplicate data. The paper's central hypothesis is that the performance of these powerful classifiers is being throttled by poor input data quality. Therefore, the main contribution is a two-part preprocessing and optimization pipeline designed to "clean" the data before classification. This pipeline consists of:

1. A systematic feature selection process to reduce redundancy.
2. A kernel modification and optimization process to tune the classifier.

The hybrid SVM-DBN model is thus presented as the beneficiary of this optimization pipeline, which is designed to unlock its true performance potential.

The "Methodology" section provides a detailed taxonomy of feature selection (FS) techniques, defining them as a process to select a subset of

relevant input variables. The objectives are to reduce computational cost, simplify the model, and, most importantly, improve performance by removing irrelevant or redundant features.

The paper categorizes FS methods as follows:

- Filter Methods: Model-agnostic, using statistical criteria.
- Wrapper Methods: Use a specific machine learning model to evaluate feature subsets.
- Embedded (Intrinsic) Methods: Feature selection is an integral part of the model's training process (e.g; Random Forest feature importance).

The authors also draw a critical distinction between feature selection (which keeps or removes original features) and dimensionality reduction (which creates new, composite features from the original ones).

The second part of the proposed pipeline is "kernel modification." The kernel function in an SVM is what enables it to handle non-linear data. The paper explicitly states its focus on the optimization of the parameters of the Radial Basis Function (RBF) kernel.

While the paper's results demonstrate that this optimization, combined with FS, leads to significant improvement, the "Methodology" section itself does not specify how this optimization is performed. It is unclear if this involves a standard grid search, a more advanced technique, or a novel modification to the RBF function itself. This "kernel modification method" remains an undefined component in the paper's description.

The paper first establishes a baseline by testing a suite of eight models on the German and Japanese credit datasets using a standard Gaussian Kernel and no specialized feature selection pipeline. The results from this baseline test showed a high variance in performance. Notably, the proposed IDCOST model, even in its baseline form, achieved a high Weighted Accuracy (94% on the German dataset, 93% on the Japanese), setting a high bar for improvement.

The paper's core empirical claim rests on the performance of these same models after applying the proposed RBF Kernel and Feature Selection pipeline. The results show a general improvement across the board for all models tested.

Result

A deeper analysis reveals a significant nuance. While the paper is titled after its IDCOST method, the data suggests that the pipeline is the true innovation, not the IDCOST model itself.

On the German dataset, the proposed IDCOST model (95% Waccuracy) does emerge as the top performer, narrowly beating XGBoost (94%) and RF (92%). However, on the Japanese dataset, the proposed IDCOST model (94%Waccuracy) is outperformed by both the DBN-based model (96%) and the Majority Voting model (95%).

This implies that the RBF+FS pipeline is a powerful, model-agnostic wrapper that benefits all models. In fact, on the Japanese dataset, the DBN-based model saw a +3 points gain (93% to 96%) from the pipeline, while the IDCOST model only gained +1 point (93% to 94%).

Conclusions

The results of this research demonstrate that a hybrid SVM and DBN approach, when augmented by a systematic feature selection and RBF kernel optimization pipeline, can significantly increase classification accuracy on imbalanced credit datasets. The paper's data empirically supports an average weighted accuracy increase of 5.75% on the German dataset and 1.875% on the Japanese dataset across a suite of eight different models. (This analysis corrects a likely typographical error in the paper, which claimed a 7.5% gain for the German dataset).

The paper's primary contribution is not the IDCOST model itself, but rather the demonstration of this RBF+FS optimization pipeline as a highly effective, model-agnostic wrapper for improving performance. The fact that the DBN based model ultimately outperformed the paper's own proposed model on the Japanese dataset is a critical finding that highlights the pipeline's broad applicability.

The author's state that a key limitation of the research is the use of benchmark case study data rather than real-world, proprietary credit data, which would be necessary for full validation.

Future research directions proposed by the authors include applying these methods to the direct prediction of credit risk, developing new scoring methods for imbalanced data, and classifying "good-behavior" versus "bad-behavior" customers. Ultimately, the paper successfully links

the economic cost of fraud to a technical solution (the FS + RBF pipeline) and the appropriate evaluation metrics (Weighted Accuracy), representing a robust, end-to-end alignment for tackling this critical FinTech challenge.

Authors' Contribution

All authors contributed to the design, implementation, and writing of this research.

Compliance with Ethical Standards

All ethical standards have been respected in this research.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

سازمان بورس و اوراق بهادار، مرکز پژوهش، توسعه و مطالعات اسلامی

فصلنامه بورس اوراق بهادار، سال هجدهم، شماره ۷۲، زمستان ۱۴۰۴، صص ۲۶۸-۲۴۱

روشی برای افزایش دقت داده‌های نامتعادل اعتباری[1]

آرش قربان‌نیادلاور[2]، صدف السادات ضیاء[3]

مقاله پژوهشی

تاریخ دریافت:۱۴۰۱/۰۸/۰۷ تاریخ بازنگری:۱۴۰۴/۰۹/۳۰

تاریخ پذیرش:۱۴۰۴/۰۹/۳۰ تاریخ انتشار:۱۴۰۴/۰۹/۳۰

نکات برجسته

- این مطالعه یک رویکرد محاسباتی ترکیبی را معرفی می‌کند که تکنیک‌های ماشین بردار پشتیبان و شبکه باور عمیق را برای طبقه‌بندی داده‌های اعتباری نامتوازن ادغام می‌کند.
- این پژوهش یک فرایند بهینه‌سازی دو مرحله‌ای را ارائه می‌دهد که شامل «انتخاب ویژگی» و «اصلاح هسته» است و هدف آن به حداقل رساندن افزونگی و تکرار داده‌ها می‌باشد.
- تمرکز اصلی بر روی معیار «دقت وزن‌دار» قرار گرفته است تا به طور مؤثر با عدم توازن کلاس‌ها و هزینه‌های نابرابر مرتبط با ارزیابی ریسک اعتباری مقابله شود.
- روش پیشنهادی منجر به افزایش کمابیش ۷/۵ درصدی در میانگین دقت وزن‌دار برای مجموعه‌داده آلمان و حدود ۲ درصد برای مجموعه‌داده ژاپن شده است.
- این پژوهش نشان می‌دهد که فرایند بهینه‌سازی (انتخاب ویژگی + هسته) به عنوان یک چارچوب قدرتمند و مستقل از مدل عمل می‌کند و عملکرد مدل‌های طبقه‌بندی مختلف را بهبود می‌بخشد.

چکیده

هدف اصلی این پژوهش ارائه روشی است که با استفاده از آن بتوانیم دقت داده‌های نامتعادل اعتباری را افزایش دهیم. کلاهبرداری مالی یک مشکل اساسی است که هم بخش مالی و هم زندگی روزمره را تحت تأثیر قرار می‌دهد و نقش مهمی در یکپارچگی و اعتماد در بخش‌های مالی و همچنین هزینه زندگی افراد دارد. طبقه‌بندی داده‌ها در ماشین بردار پشتیبان نقش بسیار مهمی در افزایش دقت دارد. به‌همین منظور ما روشی برای طبقه‌بندی داده‌های نامتعادل اعتباری با استفاده از ماشین بردار پشتیبان و شبکه باورعمیق ارائه خواهیم‌داد. با مطالعه موردی بر روش‌های شبکه باورعمیق و ماشین بردار پشتیبان داده‌های نامتعادل اعتباری مورد توجه قرار نگرفته و همچنین افزونگی داده و داده‌های تکراری نیز در این روش‌ها وجود داشته‌است. ما می‌خواهیم افزونگی داده و داده‌های تکراری را کاهش دهیم و در نهایت میزان وزن دقت داده‌ها برای داده‌های نامتعادل اعتباری را با استفاده از روش انتخاب‌ویژگی و تغییر کرنل افزایش دهیم.

واژه‌های کلیدی: داده‌های نامتعادل، داده‌های نامتعادل اعتباری، طبقه‌بندی اعتباری، دقت داده‌های نامتعادل، دقت.

طبقه‌بندی موضوعی: E5, E51, C10, C55

---

۱. doi: 10.22034/JSE.2025.12096.2066

۲. دانشیار، گروه مهندسی کامپیوتر، دانشکده فنی مهندسی، دانشگاه پیام‌نور، تهران، ایران. (نویسنده مسئول). (a_ghorbannia@pnu.ac.ir).

۳. کارشناس ارشد، گروه مهندسی کامپیوتر، دانشکده فنی مهندسی، دانشگاه پیام نور، تهران، ایران. (zia.s69@gmail.com).

استناد: قربان‌نیادلاور، آرش؛ ضیاء، صدف السادات. (۱۴۰۴). روشی برای افزایش دقت داده‌های نامتعادل اعتباری. *فصلنامه بورس اوراق بهادار*، ۱۸ (۷۲)، ۲۶۸-۲۴۱. https://10.22034/JSE.2025.12096.2066

مقدمه

با توسعه مستمر اقتصاد جهانی در سال‌های اخیر، فعالیت روزافزون شرکت‌ها در بازارهای مالی، به رشد پایدار اقتصاد ملی کمک شایانی کرده است. با این حال، عواملی چون آگاهی ناکافی از ریسک، نوسانات بازارهای مالی، کلاهبرداری‌های تجاری و سوءمدیریت، منجر به بروز شکاف در زنجیره سرمایه بسیاری از شرکت‌ها شده است. در این میان، بانک‌های کوچک به دلیل آسیب‌پذیری بیشتر در برابر ریسک‌ها، با خطر ورشکستگی روبرو هستند. این مساله به طور عمده از آنجا ناشی می‌شود که بسیاری از شرکت‌ها به جای سرمایه‌گذاری در اقتصاد واقعی، تمرکز خود را به بازارهای مالی مجازی معطوف کرده‌اند. هم‌زمان با رشد سریع تعداد درخواست‌های وام، ارزیابی ریسک اعتباری، هم برای متخصصان این حوزه و هم برای پژوهشگران، به موضوعی حیاتی تبدیل شده است. در نظام‌های ارزیابی سنتی، اطلاعات جمعیت‌شناختی و جزئیات درخواست وام، ورودی‌های اصلی برای فرآیند مهندسی ویژگی محسوب می‌شوند(ژانگ، وانگ، لو، وانگ، ما [1] ۲۰۱۷).

با ظهور عصر کلان‌داده، اگرچه مدل‌های کنترل ریسک نقش مشخصی در مدیریت اعتبار ایفا می‌کنند اما به دلیل عدم درک عمیق از داده‌ها، دقت پایینی داشته و قادر به ارزیابی جامع وام‌گیرنده نیستند. با تعمیق پژوهش‌ها در حوزه ابزارهای کلان‌داده، روش‌های نوینی در حوزه مالی به کار گرفته شده‌اند. مدل‌های سنتی کنترل ریسک با چالش‌هایی نظیر تک‌بعدی بودن و ظرفیت ارزیابی محدود روبرو هستند. در مقابل، با استفاده از روش داده‌کاوی [2] می‌توان داده‌های صنعت مالی را به‌صورت عمیق و چندبعدی تحلیل کرد و به ارزیابی جامع‌تری دست یافت. ایجاد پایگاه‌های داده توزیع‌شده و تکامل معماری‌های پلتفرم‌های داده، امکان ذخیره‌سازی و پردازش حجم وسیع‌تری از اطلاعات را فراهم آورده و ابعاد تحلیل داده را گسترش داده است. کلان‌داده می‌تواند سیستم اعتباری را بهبود بخشد و به شرکت‌ها در کاهش ریسک اعتباری کمک کند. تکنیک‌های داده‌کاوی با بهره‌گیری از داده‌های شبکه‌ای و یکپارچه‌سازی الگوریتم‌ها، قادر به بهبود مدل‌های امتیازدهی و بهینه‌سازی فرآیند اعتبارسنجی هستند و یک چرخه مدیریت پایدار و بسته را ایجاد می‌کنند. بنابراین، به‌کارگیری داده‌کاوی برای ارزیابی ریسک مالی به یکی از موضوعات کلیدی پژوهشی در این حوزه تبدیل شده است (گونارسون، واندن بروک، باسنز، اسکارسدوتیر و لومایو [3]، ۲۰۲۱).

---

1. Zhang, Wang, Lu, Wang & Ma
2. Data Mining
3. Gunnarsson, Vanden broucks, Baesens, Oskarsdottir & Lemahieu

کلاهبرداری مالی، پدیده‌ای بنیادین است که بخش مالی و زندگی روزمره را تحت تأثیر قرار داده و بر یکپارچگی و اعتماد در نظام اقتصادی و همچنین هزینه‌های معیشتی افراد اثرگذار است. این پدیده که نوعی سوءاستفاده مالی محسوب می‌شود، نگرانی عمده‌ای در جامعه اقتصادی به شمار می‌رود و خسارات چشمگیری به دولت‌ها، سازمان‌ها، شرکت‌ها و افراد تحمیل می‌کند. کلاهبرداری مالی را می‌توان عملی غیرقانونی یا نادرست تعریف کرد که از طریق روش‌های غیراخلاقی به نفع یک فرد یا سازمان تمام می‌شود. تکنیک‌های کشف تقلب با هدف شناسایی فعالیت‌های غیرعادی در تراکنش‌های گذشته طراحی شده‌اند تا از اقداماتی جلوگیری کنند که در آن کلاهبرداران در صدد تضییع ارزش‌های ایجاد شده توسط سازمان‌ها هستند. بخشی از چالش کنونی بانک‌ها و مؤسسات مالی، افزایش حجم مطالبات معوق و مشکوک‌الوصول است که ریشه در عدم بهره‌گیری از یک نظام کارآمد برای اندازه‌گیری و مدیریت ریسک اعتباری دارد (ملک محمدی، سعیدی و متین فرد ۱۳۹۹).

اگرچه روش‌های گوناگونی برای تشخیص تقلب، از جمله رتبه‌بندی اعتباری مؤسسات، ارائه شده است (محقق‌نیا، قربانی‌زاده و خان‌زاده ۱۴۰۰)، اما این رویکردها در برابر تکامل مداوم شیوه‌های کلاهبرداری و ظهور فناوری‌های نوین مانند ارزهای دیجیتال، کارایی خود را از دست می‌دهند. هر سیستم تجارت الکترونیکی که تراکنش‌های آنلاین، به‌ویژه خدمات مالی را پردازش می‌کند، در معرض حملات کلاهبرداران قرار دارد. بنابراین، مبارزه با تقلب به یک حوزه پژوهشی جذاب برای پژوهشگران تبدیل شده است. اهمیت این مسئله، پژوهشگران را به توسعه روش‌های پیشرفته‌تر برای کشف تقلب و تخمین ریسک آن سوق داده است.

در این پژوهش، با به‌کارگیری یک رویکرد ترکیبی مبتنی بر ماشین بردار پشتیبان[1] و شبکه باور عمیق[2]، به مساله افزونگی داده پرداخته شده است. افزون بر این، از طریق ارزش‌گذاری داده‌های نامتعادل[3] و طبقه‌بندی دقیق ورودی‌ها، دقت مدل افزایش یافته است. در نهایت با یکپارچه‌سازی پارامترها در تابع هدف و شاخص‌گذاری ویژگی‌ها[4]، موفق به بهبود معیار دقت وزنی شده‌ایم.

---

1. Support Vector Machine
2. Deep Belief Network
3. Imbalanced data
4. Properties

مبانی نظری و توسعه فرضیه‌ها

مساله عدم تعادل داده‌ها در وظایف طبقه‌بندی[1] چالشی شناخته‌شده است که برای حل آن، راه‌حل‌های متعددی ارائه شده است. این رویکردها را می‌توان در چهار دسته اصلی طبقه‌بندی کرد: روش‌های سطح داده، اصلاحات سطح الگوریتم، روش‌های مبتنی بر هزینه[2]و رویکردهای ترکیبی[3]. هر یک از این روش‌ها مزایا و معایب خاص خود را دارند و تاکنون راه‌حل جامعی که برای تمام سناریوها بهینه باشد، ارائه نشده است.

روش‌های سطح داده، مانند نمونه‌گیری مجدد[4] و بیش‌نمونه‌برداری[5]، با تغییر توزیع داده‌ها به دنبال ایجاد تعادل هستند. رویکردهای سطح الگوریتم، با اصلاح الگوریتم‌های طبقه‌بندی، حساسیت آن‌ها را نسبت به کلاس اقلیت (کلاس کمتر اما مهم‌تر) افزایش می‌دهند. روش‌های مبتنی بر هزینه، با تخصیص وزن بیشتر به کلاس اقلیت یا جریمه بالاتر برای طبقه‌بندی اشتباه آن، مدل را هدایت می‌کنند. در نهایت روش‌های ترکیبی که بتازگی محبوبیت یافته‌اند، با تجمیع چندین طبقه‌بندی‌کننده[6] ضعیف، یک مدل قدرتمند و جامع ایجاد می‌کنند. در حوزه طبقه‌بندی داده‌های اعتباری نامتعادل، پژوهشگران متعددی با هدف افزایش دقت، مدل‌های گوناگونی را پیشنهاد کرده‌اند.

روش‌های سنتی ارزیابی اعتبار، بیشتر بر الگوریتم‌های یادگیری ماشین مانند ماشین بردار پشتیبان، پرسپترون چندلایه[7] و به‌ویژه درخت تصمیم[8] استوار بوده‌اند(تیان، شیائو، فنگ و وی[9] ۲۰۲۰). پژوهش‌های جدیدتر، کارایی روش‌های پیشرفته‌تری مانند درخت تصمیم تقویت‌شده با گرادیان[10] را به نمایش گذاشته‌اند(چنگ، چنگ و وو[11] ۲۰۱۸). در این مطالعات، عملکرد مدل‌ها معمولاً پس از اعمال مراحل پیش‌پردازش[12] و انتخاب ویژگی[13] ارزیابی و مقایسه می‌شود(ژانگ، وانگ، چونگ و وانگ[14] ۲۰۲۱).

---

1. Classification
2. Cost-based Methods
3. Ensemble methods
4. Resampling
5. Over sampling
6. Classifiers
7. Multilayer perceptron(MLP)
8. Decision tree
9. Tian, Xiao, Feng & Wei
10. Gradient Boosting Decision Tree
11. Chang, Chang & Wu
12 Preprocessing
13 Feature Selection
14 Zhang, Wang, Chung & Wang

درخت تصمیم، یک مدل یادگیری ماشین[1] است که با طرح پرسش‌های متوالی، داده‌ها را به‌صورت بازگشتی به زیرمجموعه‌های کوچک‌تر تقسیم می‌کند تا به یک نتیجه یا برچسب[2] دست یابد. ساختار این مدل، مشابه یک فلوچارت درختی است که در آن هر گره داخلی نمایانگر یک ویژگی، هر شاخه نشان‌دهنده یک قاعده تصمیم‌گیری و هر گره برگ، خروجی نهایی (برچسب کلاس) را مشخص می‌کند. شکل ۱ قسمتی از تصمیم گیری را نشان می‌دهد.

شکل ۱. نمایش نمونه‌ای از تصمیم‌گیری در درخت تصمیم

درخت تصمیم، مدلی با ساختار سلسله‌مراتبی و مشابه فلوچارت است. در این ساختار، هر گره[3] داخلی نمایانگر یک ویژگی (متغیر) از مجموعه‌داده[4]، هر شاخه[5] نشان‌دهنده یک قاعده تصمیم‌گیری، و هر گره برگ[6] معرف خروجی یا نتیجه نهایی است. این مدل با شروع از گره ریشه[7]، داده‌ها را بر اساس مقادیر ویژگی‌ها به‌صورت بازگشتی[8] به زیرمجموعه‌های کوچک‌تر تقسیم می‌کند، فرآیندی که به آن «پارتیشن‌بندی بازگشتی» گفته می‌شود. ساختار بصری و شفاف درخت تصمیم، فرآیند تصمیم‌گیری انسان را شبیه‌سازی می‌کند و به همین دلیل، تفسیر و درک آن بسیار ساده است.

---

1. Machine Learning
2. Label
3. Node
4. Dataset
5. Branch
6. Leaf
7. Root
8. Recursive

با این حال، یکی از چالش‌های اصلی درختان تصمیم، تمایل آن‌ها به «بیش‌برازش[1]» است. این پدیده باعث می‌شود مدل بر روی مجموعه‌داده اعتبارسنجی[2] عملکرد مطلوبی داشته باشد، اما در روبرویی با مجموعه‌داده آزمایشی[3] جدید، دقت آن به‌شدت افت کند. برای چیرگی بر این محدودیت، روش‌های «یادگیری گروهی[4]» توسعه یافته‌اند.

پژوهش‌های پیشین نیز با استفاده از الگوریتم‌هایی مانند ماشین بردار پشتیبان(داناس و گارسوا [5]۲۰۱۲؛ ژانگ، ژانگ، شو و هائو[6] ۲۰۱۸) و شبکه باور عمیق (یو، ژو، تانگ و چن [7]۲۰۱۸) به نتایج چشمگیری در این حوزه دست یافته‌اند. پژوهش حاضر در ادامه این تلاش‌ها، با هدف دستیابی به دقت بهینه‌تر، بر بهینه‌سازی پارامترهای کرنل تمرکز دارد. افزون بر این، با توجه به اینکه هزینه ناشی از یک طبقه‌بندی اشتباه (مانند تأیید یک مشتری متقلب) بسیار بیشتر از سود حاصل از یک مشتری خوب است، «ماتریس درآمد» به‌عنوان ابزاری برای در نظر گرفتن هزینه‌های نامتقارن معرفی می‌شود که در جدول ۱ نمایش داده شده است.

جدول ۱. ماتریس درآمد

	مشتری خوب (پیش بینی)	مشتری بد (پیش بینی)
مشتری خوب (واقعی)	TN(Interest)	FP(0)
مشتری بد (واقعی)	FN(0)	TP(Principal)

باتوجه به جدول ۱، روابط زیر را برای محاسبه دقت داریم:

$$\text{TPaccuracy} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (۱)$$

1. Overfitting
2. Validation Dataset
3. Test Dataset
4. Ensemble Learning
5. Danenas & Garsva
6. Zhang, Zhang, Xu & Hao
7. Yu, Zhou, Tang & Chen

نتیجه دقت درست مثبت[1] از درست مثبت[2] در صورت کسر و جمع درست مثبت به همراه نادرست منفی[3] در مخرج کسر بدست می آید.

$$\text{TNaccuracy} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad (۲)$$

نتیجه دقت درست منفی[4] از درست منفی در صورت کسر و جمع درست منفی[5] به همراه نادرست مثبت[6] در مخرج کسر بدست می آید.

$$\text{Waccuracy} = \frac{\text{Principal} \times \text{TPaccuracy} + \text{Interest} \times \text{IR} \times \text{TNaccuracy}}{\text{Principal} + \text{Interest} \times \text{IR}} \qquad (۳)$$

IR نسبت عدم‌تعادل[7] مجموعه است، نسبت داده‌های مثبت به داده‌های منفی یا نسبت مشتری‌های خوب به مشتری‌های بد(یو و همکارانش ۲۰۱۸). بطور مثال برای مجموعه‌داده اعتباری آلمان داده‌های مثبت ۷۰۰ و داده‌های منفی ۳۰۰ است و نسبت عدم‌تعادل این مجموعه‌داده برابر ۲/۳ است.

جدول ۲ نتیجه مجموعه‌داده آلمانی و جدول۳ نتیجه مجموعه‌داده ژاپنی را نشان می‌دهد. نمودار ۱ تا ۷ نتایج این دو جدول را نمایش می‌دهد.

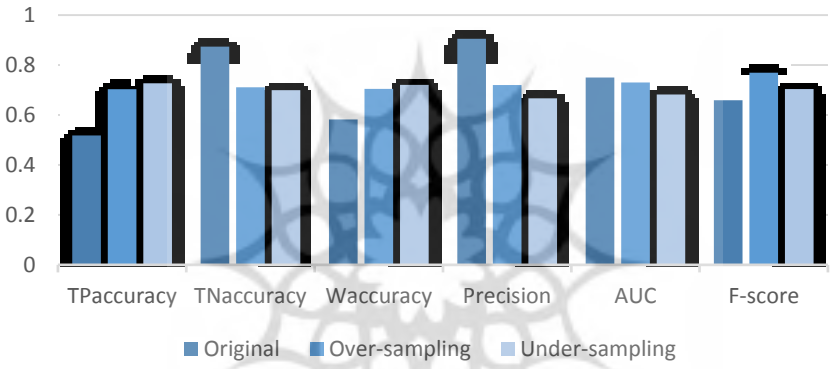
جدول۲. مجموعه‌داده آلمانی با کرنل گاوسی

Model	Data pre-processing	TPaccuracy	TNaccuracy	Waccuracy	Precision	AUC	F-score
Single SVM	Original	۵۱/٪.۸۰	۸۷/٪.۳۷	۵۸/٪.۲۱	۹۰/٪.۵	٪.۷۵	۶۵/٪.۹
	Over-sampling	۷۰/٪.۳۰	۷۱/٪.۱۰	۷۰/٪.۴۴	٪.۷۲	٪.۷۳	۷۶/٪.۹
	Under-sampling	۷۲/٪.۷۰	۶۹/٪.۹۳	۷۲/٪.۰۶	۶۶/٪.۶۵	۶۸/٪.۲۰	۷۰/٪.۴
Majority voting	Over-sampling	۸۰/٪.۰۰	۶۸/٪.۴۷	۷۷/٪.۹۲	۷۷/٪.۵	۶۹/٪.۴۸	۶۸/٪.۳
DBN-based	Over-sampling	۸۳/٪.۶۰	۶۴/٪.۶۳	۸۰/٪.۱۸	۸۰/٪.۲۵	٪.۸۳	۸۴/٪.۱
XGBoost	Over-sampling	۴۷/٪.۸	۸۵/٪.۴	۸۹/٪.۶۲	۸۳/٪.۴	۶۵/٪.۶۷	۵۰/٪.۲
RF	Over-sampling	٪.۸۶	۸۶/٪.۶۵	۹۰/٪.۲	۸۵/٪.۲	۵۹/٪.۶	۴۹/٪.۶
IDCOST	Over-sampling	۸۱/٪.۷	۹۲/٪.۴	۹۴/٪.۶	۹۲/٪.۳۵	۷۳/٪.۳	۶۳/٪.۳

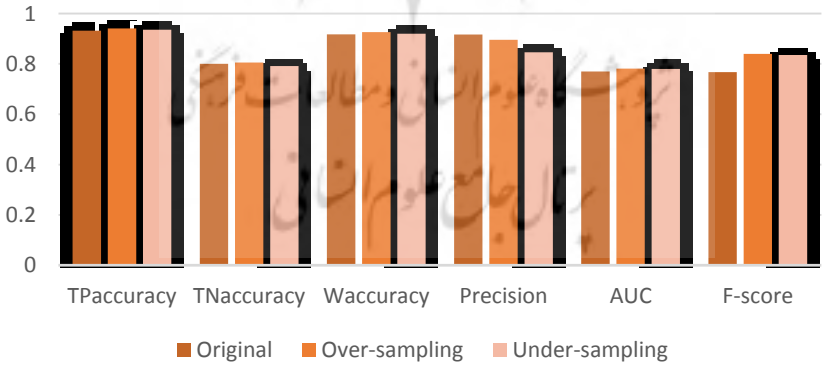
1. True Positive Accuracy, TPaccuracy
2. True Positive, TP
3. False Negative, FN
4. False negative accuracy, TNaccuracy
5. True negative, TN
6. False positive, FP
7. Imbalanced ratio, IR

جدول۳. مجموعه‌داده ژاپنی با کرنل گاوسی

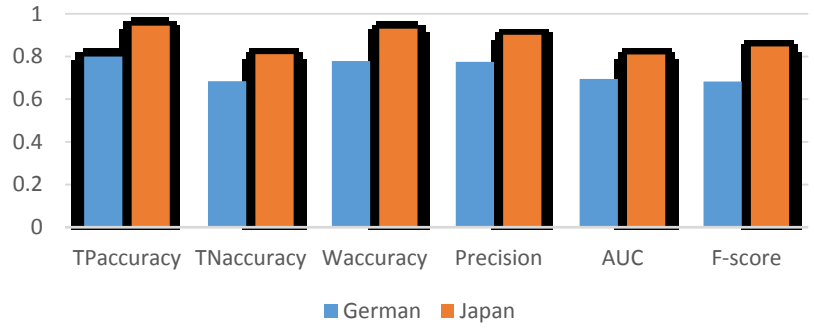
Model	Data pre-processing	TPaccuracy	TNaccuracy	Waccuracy	Precision	AUC	F-score
Single SVM	Original	۹۳//.۱۷	۸۰//.۰۰	۹۱//.۸۴	۹۱//.۷	٪۷۷	۷۶//.۷
	Over-sampling	۹۴//.۰۸	۸۰//.۵۷	۹۲//.۶۲	۸۹//.۵۶	۷۸//.۱۲	٪۸۴
	Under-sampling	۹۳//.۵۸	۷۹//.۱۷	۹۱//.۹۸	۸۴//.۶	٪۷۸	۸۳//.۵
Majority voting	Over-sampling	۹۴//.۵۰	۸۱//.۱۴	۹۳//.۰۶	۹۰//.۲	٪۸۱	۸۴//.۷
DBN-based	Over-sampling	۹۴//.۵۰	۸۱//.۱۴	۹۳//.۰۶	۹۲//.۴۱	۹۰//.۴	۹۱//.۸
XGBoost	Over-sampling	۸۳//.۶۲	۸۴//.۸	۸۴//.۹	۸۳//.۶	۸۴//.۲۴	٪۸۲
RF	Over-sampling	۸۲//.۳۶	۸۵//.۶۸	۸۶//.۶	۸۴//.۷	۸۳//.۷	۸۱//.۵
IDCOST	Over-sampling	۹۱//.۷۴	۸۹//.۴۵	۹۳//.۲	۹۰//.۶	۹۰//.۲	۸۷//.۲



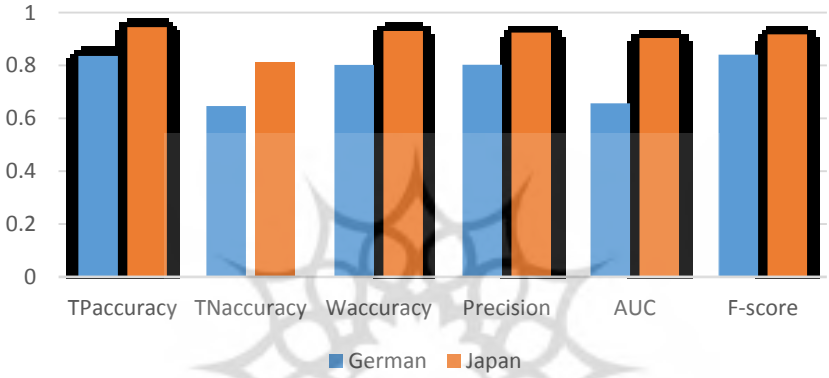
نمودار ۱. Single SVM داده آلمان



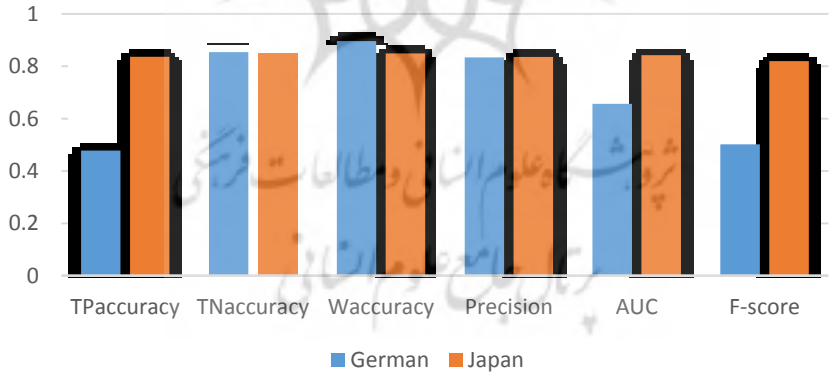
نمودار۲. Single SVM داده ژاپن



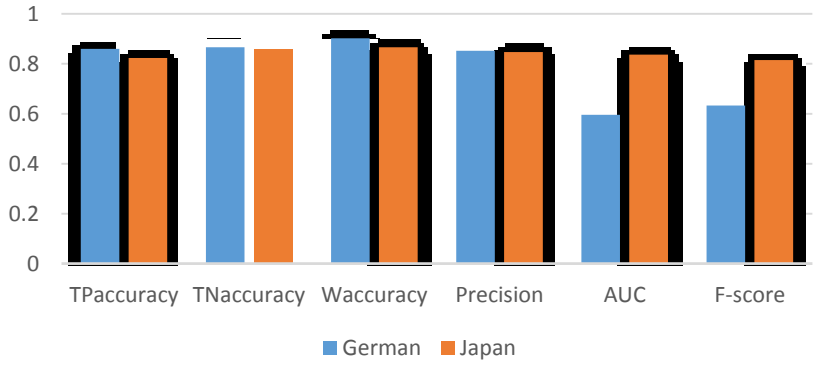
نمودار۳. Majority voting آلمان و ژاپن



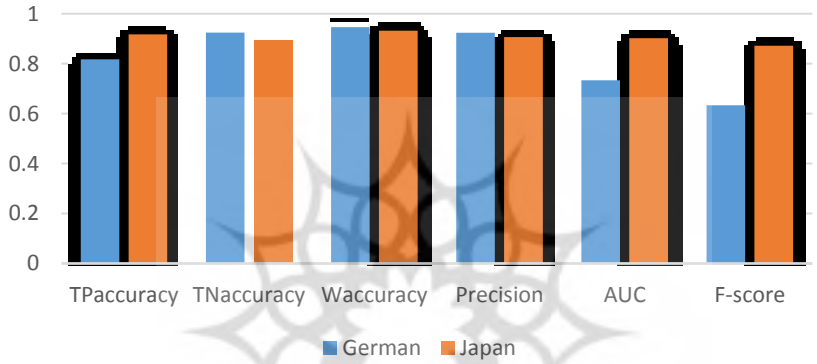
نمودار٤. DBN-based آلمان و ژاپن



نمودار٥. XGBoost آلمان و ژاپن



نمودار٦. Random Forest آلمان و ژاپن



نمودار۷. IDCOST آلمان و ژاپن

درجدول ۴ ابزار تحلیل مورداستفاده و نتایج بدست آمده در برخی پژوهش‌ها را نشان داده‌ایم.

جدول ٤. جدول پژوهش‌ها

نام نویسندگان	سال	عنوان	مجموعه‌داده	الگوریتم مورد استفاده	نتایج
چنگ و همکارانش	۲۰۱۸	کاربرد درخت‌های تقویت گرادیان شدید در ساخت مدل‌های ارزیابی ریسک اعتباری	داده‌های اعتباری نامتعادل	XGBoost (درخت تقویت گرادیان شدید)، SVM، رگرسیون لجستیک	الگوریتم XGBoost نتایج بهتری نسبت به سایر طبقه‌بندها از نظر AUC و دقت کسب کرد.
داناس و همکارانش	۲۰۱۲	مدل‌سازی ارزیابی ریسک اعتباری با استفاده از طبقه‌بندهای SVM خطی تکاملی	داده‌های واقعی ورشکستگی	SVM خطی با بهینه‌سازی ازدحام ذرات (PSO)	مدل پیشنهادی نتایج امیدوارکننده‌ای را بر روی داده‌های واقعی ورشکستگی نشان داد.

نام نویسندگان	سال	عنوان	مجموعه‌داده	الگوریتم مورد استفاده	نتایج
داس، مگاناث، بهرا، ممتاز، الکویری و فاروق[1]	۲۰۲۴	QFDNN: یک شبکه عصبی عمیق با ویژگی‌های کوانتومی برای کشف تقلب و پیش‌بینی وام	داده‌های کشف تقلب کارت اعتباری و پیش‌بینی صلاحیت وام	شبکه عصبی عمیق با ویژگی‌های کوانتومی (QFDNN)	دستیابی به دقت رقابتی (۸۲/۲٪ برای کشف تقلب) با سربار محاسباتی کاهش‌یافته.
دینگ، جیا، ژوانگ و دینگ[2]	۲۰۲۲	رگرسیون نامتعادل عمیق با استفاده از یادگیری حساس به هزینه و انتقال ویژگی عمیق	داده‌های تجربی یاتاقان (برای تخمین عمر مفید)	چارچوب CSL-DFT (یادگیری حساس به هزینه و انتقال ویژگی عمیق)	کاهش قابل توجه خطای میانگین مربعات (RMSE) و پیشی گرفتن از سایر روش‌ها در دقت پیش‌بینی.
قربان‌نیا دلاور و ضیاء	۲۰۲۵	IDCOST: روشی برای افزایش سرویس معیار داده با امتیازدهی داده‌های نامتعادل اعتباری	داده‌های نامتعادل اعتباری	متد IDCOST مبتنی بر SVM و انتخاب ویژگی	دستیابی به حساسیت بالاتر در کشف تقلب و مقایسه با سایر روش‌ها.
جیوشتی، گوارنرا، کاسو و باتیاتو[3]	۲۰۲۵	تقلب فقط یک پدیده نادر نیست: یک رویکرد توجه نمونه اولیه علّی برای پیش‌نمونه‌برداری واقعی	داده‌های تراکنش کارت اعتباری	طبقه‌بند توجه نمونه اولیه علّی (CPAC) با VAE-GAN	دستیابی به عملکرد برتر با F1-score معادل ۹۳/۱۴٪ و Recall معادل ۹۰/۱۸٪
گونارسون و همکارانش	۲۰۲۱	یادگیری عمیق برای امتیازدهی اعتباری: انجام دهیم یا ندهیم؟	مجموعه‌داده‌های اعتباری آلمان، استرالیا، ژاپن و تایوان	یادگیری عمیق (MLP, DBN) در مقایسه با XGBoost	XGBoost بهترین عملکرد را داشت؛ شبکه‌های عصبی عمیق از همتایان کم‌عمق خود بهتر عمل نکردند.
گوپتا، حسن، ماجهی، پروین، زمانی، آنیتا، اوجا، سینگ و مودولی[4]	۲۰۲۵	چارچوب بهبودیافته برای کشف تقلب کارت اعتباری با استفاده از انتخاب ویژگی و مدل گروهی پشته‌ای	پنج مجموعه‌داده متنوع با نسبت‌های عدم تعادل مختلف	مدل گروهی پشته‌ای (Stacking Ensemble) با انتخاب ویژگی ترکیبی	روش پیشنهادی در تمام مجموعه‌داده‌ها عملکرد بهتری نسبت به رویکردهای پایه موجود داشت.
شارما[5]	۲۰۲۵	شبکه‌های عصبی مبتنی بر هوش مصنوعی برای تشخیص الگوهای ناهنجار در تراکنش‌های مالی آنی	داده‌های تراکنش مالی آنی	شبکه‌های عصبی (CNNs, RNNs/LSTMs) برای تشخیص ناهنجاری	سیستم‌های مبتنی بر هوش مصنوعی نرخ کشف تقلب را افزایش داده و هشدارهای کاذب را کاهش می‌دهند.
تیان و همکارانش	۲۰۲۰	ارزیابی ریسک اعتباری مبتنی بر درخت تصمیم تقویت گرادیان	داده‌های اعتباری نامتعادل	درخت تصمیم تقویت گرادیان (GBDT) با SMOTE	مدل GBDT به بالاترین دقت (۹۲/۱۹٪)، F1-score (۹۱/۸۳٪) و AUC (۰/۹۷) دست یافت.
یو، یائو، وانگ و لای[6]	۲۰۱۱	ارزیابی ریسک اعتباری با استفاده از طبقه‌بند SVM حداقل مربعات وزنی	دو مجموعه‌داده اعتباری عمومی	SVM حداقل مربعات وزنی (LSSVM) با طراحی آزمایش (DOE)	طبقه‌بند LSSVM وزنی پیشنهادی نتایج طبقه‌بندی امیدوارکننده‌ای تولید کرد.

1. Das, Meghanath, Behera, Mumtaz, Al-Kuwari & Farouk
2. Ding, Jia, Zhuang & Ding
3. Giusti, Guarnera, Casu & Battiato
4. Gupta, Hassan, Majhi, Parveen, Zamani, Anitha, Ojha, Singh & Muduli
5. Sharma
6. Yu, Yao, Wang & Lai

نام نویسندگان	سال	عنوان	مجموعه‌داده	الگوریتم مورد استفاده	نتایج
یو و همکارانش	۲۰۱۸	یک پارادایم یادگیری گروهی مبتنی بر DBN و نمونه‌برداری مجدد SVM برای طبقه‌بندی اعتباری	داده‌های اعتباری نامتعادل (آلمان و ژاپن)	شبکه باور عمیق (DBN) به عنوان مدل گروهی برای SVM	عملکرد طبقه‌بندی به ویژه برای داده‌های نامتعادل به طور مؤثری بهبود یافت.
ژانگ و همکارانش	۲۰۱۷	یک الگوریتم بهبودیافته SMO برای ارزیابی ریسک اعتباری مالی	داده‌های اعتباری چین (CBRC) و دو مجموعه‌داده UCI	الگوریتم بهینه‌سازی حداقل متوالی بهبودیافته (FV-SMO) برای SVM	الگوریتم FV-SMO هزینه محاسباتی را کاهش داد و از پنج روش دیگر در دقت پیشی گرفت.
ژانگ و همکارانش	۲۰۱۸	یادگیری چندنمونه‌ای برای ارزیابی ریسک اعتباری با داده‌های تراکنش	پنج مجموعه‌داده واقعی از دو بانک تجاری بزرگ در چین	یادگیری چندنمونه‌ای (MIL) با تابع پایه شعاعی (RBF)	ترکیب داده‌های رفتاری پویا با اطلاعات ایستا، عملکرد پیش‌بینی را به طور قابل توجهی بهبود بخشید.
ژانگ و همکارانش	۲۰۲۱	ماشین‌های بردار پشتیبان با دانش پیشین تکامل ویژگی	-	KFEP-اصلاح‌شده SVM (SVM)	مدل پیشنهادی توانایی تعمیم‌دهی بالاتری نسبت به SVM کلاسیک نشان داد.
قربان‌نیا دلاور و جعفری	۲۰۱۵	یک روش برای کاهش طبقه‌بندی داده با استفاده از تکنیک وزن‌دهی در SVM+	-	SVM+	یک روش بهینه‌سازی شده برای الگوریتم SVM تحت عنوان **DCSVM+** پیشنهاد شده است که با کاهش داده‌های تکراری و استفاده از تکنیک وزن‌دهی، زمان پردازش را کاهش داده و دقت طبقه‌بندی را بهبود می‌بخشد.
قربان‌نیا دلاور، نوری لاهرود و ذکریاپناه گشتی	۲۰۱۱	ERPDRT: یک چارچوب جدید بلادرنگ در منابع سیستم توزیع‌شده یکپارچه، امن با مکانیسم‌های داده کاوی	-	چارچوب جدید ERPDRT ارائه شده است.	یک چارچوب جهت استفاده از زمان واقعی در سیستم‌های توزیع‌شده با امنیت و داده کاوی ارائه شده که با استفاده از عوامل ارزیابی شده، قابلیت هم‌زمانی و بهینه‌سازی زمان پاسخگویی به مشتریان را بهبود می‌بخشد. همچنین، استفاده از الگوریتم‌های داده کاوی جدید باعث افزایش بهره‌وری و رضایت مشتریان نسبت به روش‌های قبلی می‌شود

روش‌شناسی پژوهش

روش‌شناسی این پژوهش بر پایه رویکردی تحلیلی و کمی استوار است. در این پژوهش، از مجموعه‌داده‌های اعتباری نامتعادل آلمان[1] و ژاپن[2] استفاده شده که هر دو مجموعه‌داده از مخزن داده UCI هستند و تمامی مراحل پیاده‌سازی و تحلیل مدل‌ها در محیط برنامه‌نویسی پایتون و با استفاده از نوت‌بوک‌های ژوپیتر انجام گرفته‌است.

1. http://archive.ics.uci.edu/dataset/144/statlog+german+credit+data
2. http://archive.ics.uci.edu/ml/datasets/Credit+Approval

مجموعه‌داده آلمان یکی از مشهورترین و قدیمی‌ترین مجموعه‌داده‌ها در حوزه یادگیری ماشین است این مجموعه‌داده به عنوان یک معیار استاندارد برای سنجش الگوریتم‌های طبقه‌بندی استفاده می‌شود. هدف اصلی این مجموعه‌داده ارزیابی ریسک اعتباری است. این مجموعه‌داده شامل ۱۰۰۰ نمونه و دارای ۲۰ ویژگی است به همراه یک ویژگی هدف که مشخص‌کننده مشتری خوش‌حساب یا بدحساب است، ۷۰۰ نمونه مشتری خوش‌حساب، ۳۰۰ نمونه مشتری بدحساب دارد و این مجموعه‌داده ترکیبی از داده‌هایی با نوع عددی و متنی است. همچنین نسبت عدم‌تعادل این مجموعه‌داده ۲/۳۳ است یعنی به ازای هر ۲/۳۳ مشتری خوش‌حساب، یک مشتری بدحساب وجود دارد.

چالش‌های این مجموعه‌داده عبارتند از: عدم توازن کلاس‌ها، ماتریس هزینه برای کاهش ریسک ضرر مالی و وجود داده‌های متنی یا کدگذاری شده که نیازمند پیش‌پردازش هستند.

مجموعه‌داده ژاپنی یکی دیگر از مخزن داده‌های UCI که شامل ۶۹۰ نمونه و دارای ۱۵ ویژگی است به همراه یک ویژگی که مشخص می‌کند نمونه مثبت یا منفی است، ۳۰۷ نمونه مثبت (تائید شده) و ۳۸۳ نمونه منفی (رد شده) دارد، این مجموعه‌داده همانند داده‌های قبلی ترکیبی از انواع داده‌ها است. نسبت عدم تعادل این مجموعه‌داده ۱/۲۵ است یعنی به ازای ۱/۲۵ نمونه منفی یک نمونه مثبت وجود دارد. چالش این مجموعه‌داده عدم توازن کلاس‌ها و داده‌های گمشده است.

ویژگی منحصر به‌فرد این مجموعه‌داده محرمانگی کامل داده‌ها است، تمام نام ستون‌ها و مقادیر آنها پنهان‌سازی شده‌اند. اگرچه نام ستون‌ها حذف شده‌اند اما براساس تحلیل‌های آماری و مقایسه با داده‌های واقعی، حدس‌هایی درباره ماهیت ستون‌ها زده شده‌است.

با درنظر گرفتن مجموعه‌داده‌های ژاپنی و آلمانی، انتخاب صحیح طبقه‌بندی درکنار انتخاب‌ویژگی و همچنین استفاده از الگوریتم DBN، منجر به بهبود جریان کاری نسبت به سایر روش‌های مورد مطالعه شده است، درواقع انتخاب‌ویژگی باعث توازن بهتر داده‌ها و همچنین دسترسی سریع‌تر شده‌است.

رویکرد اصلی این مقاله بر یک مدل ترکیبی استوار بر ماشین بردار پشتیبان، شبکه باور عمیق و الگوریتم بسته‌بندی[1] بنا شده است. برای ارتقای عملکرد این مدل، دو تکنیک کلیدی به کار گرفته شد: بهینه‌سازی پارامترهای کرنل تابع پایه شعاعی[2] و اعمال یک فرآیند نظام‌مند انتخاب ویژگی. هدف از این ترکیب، کاهش افزونگی داده و افزایش دقت پیش‌بینی مدل است.

---

۱. Bagging

۲. Radial Basis Function (RBF)

انتخاب ویژگی، فرآیندی است که طی آن زیرمجموعه‌ای از مرتبط‌ترین متغیرهای ورودی برای ساخت یک مدل پیش‌بینی انتخاب می‌شوند(یو و همکارانش ۲۰۱۱). هدف اصلی این فرآیند، کاهش هزینه محاسباتی، ساده‌سازی مدل و بهبود عملکرد آن از طریق حذف ویژگی‌های نامرتبط یا اضافی است. روش‌های انتخاب ویژگی را می‌توان به دسته‌های مختلفی تقسیم کرد:

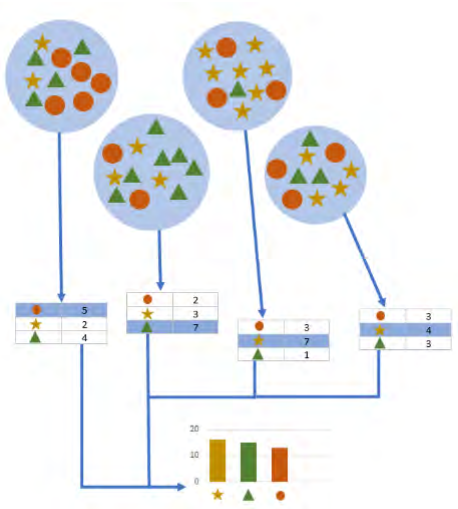
- روش‌های فیلتر[۱]: این روش‌ها ویژگی‌ها را بر اساس معیارهای آماری و مستقل از الگوریتم یادگیری ارزیابی می‌کنند. اگرچه این روش‌ها سریع و کارآمد هستند، اما انتخاب معیار آماری مناسب، بسته به نوع داده‌ها، می‌تواند چالش‌برانگیز باشد.
- روش‌های بسته‌بندی[۲]: در این رویکرد، زیرمجموعه‌های مختلفی از ویژگی‌ها با استفاده از یک مدل یادگیری ماشین ارزیابی شده و ترکیبی که بهترین عملکرد را بر اساس یک معیار مشخص دارد، انتخاب می‌شود. این روش‌ها معمولاً دقت بالاتری نسبت به روش‌های فیلتر دارند اما از نظر محاسباتی سنگین‌تر هستند.
- روش‌های ذاتی[۳]: در این دسته، فرآیند انتخاب ویژگی بخشی از خود الگوریتم یادگیری است. مدل‌هایی مانند رگرسیون لاسو[۴] و جنگل تصادفی به‌طور خودکار ویژگی‌های مهم را در حین فرآیند آموزش شناسایی و انتخاب می‌کنند.

همچنین، می‌توان روش‌های انتخاب ویژگی را بر اساس استفاده از متغیر هدف، به دو دسته با نظارت[۵] و بدون نظارت[۶] تقسیم کرد. روش‌های با نظارت از متغیر هدف برای ارزیابی و انتخاب ویژگی‌ها استفاده می‌کنند، در حالی که روش‌های بدون نظارت، متغیر هدف را نادیده گرفته و بر اساس ویژگی‌های درونی داده‌ها (مانند همبستگی) عمل می‌کنند.

باید توجه داشت که انتخاب ویژگی با کاهش ابعاد تفاوت دارد. انتخاب ویژگی، متغیرهای اصلی را حفظ یا حذف می‌کند، در حالی که کاهش ابعاد، ویژگی‌های جدیدی را از ترکیب خطی یا غیرخطی متغیرهای اصلی ایجاد می‌نماید. شکل ۲، نمای کلی فرآیند انتخاب ویژگی را نشان می‌دهد.

---

۱. Filter Methods
۲. Wrapper Methods
۳. Embedded/Intrinsic Methods
۴. Lasso
۵. Supervised
۶. Unsupervised



شکل ۱. انتخاب‌ویژگی

یافته‌های پژوهش

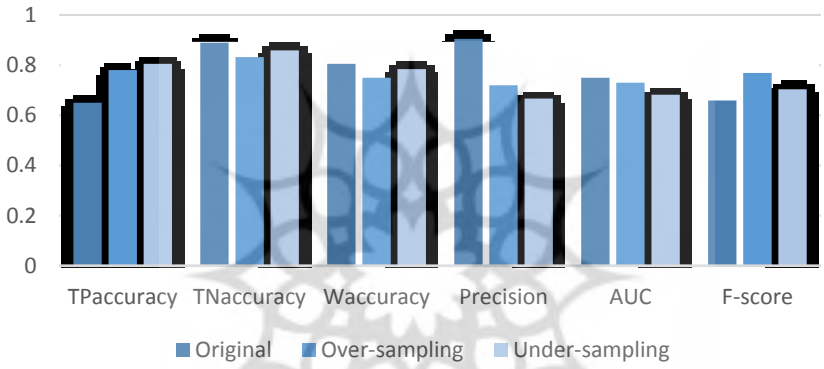
در جدول های ۵ و۶ نشان داده شده‌است که مجموعه‌داده‌ها را برای شش مدل اعمال کردیم. نمودارهای ۸ تا ۱۴ نتایج جدول های ۵ و ۶ را به روشنی نشان می‌دهد.

جدول۵. نتیجه مجموعه‌داده آلمانی با کرنل RBF و روش انتخاب‌ویژگی

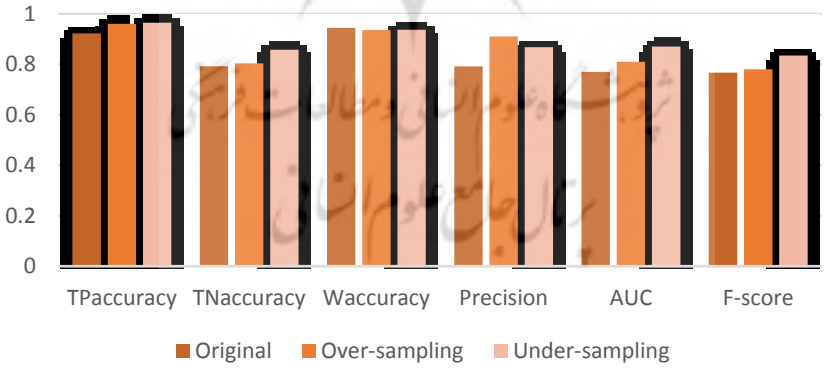
Model	Data pre-processing	TPaccuracy	TNaccuracy	Waccuracy	Precision	AUC	F-score
Single SVM	Original	۶۵/۰۰	۸۸/۹۲	۶۵/۰۹	۹۲/۳	٪۷۷	۶۸/۵
	Over-sampling	۷۸/۰۶	۸۳/۲۰	۸۵/۹۰	۷۶/۴	٪۷۳	۷۶/۹
	Under-sampling	۸۰/۵۴	۷۵/۰۰	۷۸/۴۴	۶۸/۵	۶۹/۶	۷۲/۶
Majority voting	Over-sampling	۷۸/۸۱	۸۳/۷۲	۷۹/۶۹	۷۶/۵	۷۱/۸۸	۶۳/۹
DBN-based	Over-sampling	۸۷/۳۵	۹۲/۹۸	۸۸/۱۶	۸۶/۳۴	۸۵/۸	۸۶/۶۲
XGBoost	Over-sampling	۵۰/۳	۸۸/۶	۹۴/۶	۸۴/۶	۶۶/۶۷	۵۲/۴
RF	Over-sampling	٪۸۷	۸۹/۲	۹۲/۴٪	۸۸/۶	۶۰/۹	۵۱/۸
IDCOST	Over-sampling	۸۳/۷	۹۴/۴	۹۵/۶	۹۴/۱	۷۹/۹	۶۵/۴

جدول۶. نتیجه مجموعه‌داده ژاپنی براساس کرنل RBF و روش انتخاب‌ویژگی

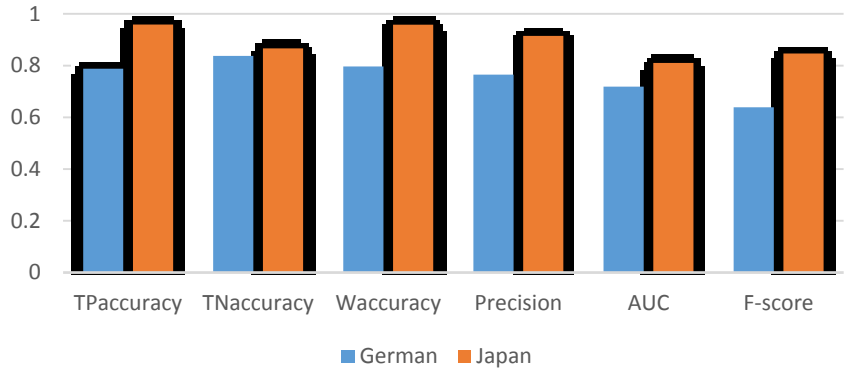
Model	Data pre-processing	TPaccuracy	TNaccuracy	Waccuracy	Precision	AUC	F-score
Single SVM	Original	۹۵/۰۲۵	۷۹/۰۲۲	۹۴/۰۴۴	۹۳/۰۲	۷۸/۵	۷۴/۰۹
	Over-sampling	۹۶/۰۰۵	۸۰/۰۳۹	۹۳/۰۵۲	٪۹۱	٪۸۱	٪۸۵
	Under-sampling	۹۶/۰۴۸	۸۵/۰۷۸	۹۳/۰۶۶	۸۶/۰۸	٪۷۸	۸۷/۰۸
Majority voting	Over-sampling	۹۵/۰۸۸	۸۶/۰۶۵	۹۵/۰۸۱	۹۱/۰۵	٪۸۱	۸۶/۰۷
DBN-based	Over-sampling	۹۷/۰۳۰	۸۸/۰۵۳	۹۶/۰۰۰	۹۴/۰۵	۹۲/۰۳	۹۴/۰۸
XGBoost	Over-sampling	۸۵/۰۷	۸۶/۰۸	۸۶/۰۳	۸۵/۰۸	۸۶/۰۳	٪۸۵
RF	Over-sampling	۸۴/۰۱	۸۶/۰۱	۸۷/۰۹	۸۵/۰۳	۸۵/۰۱	۸۳/۰۵
IDCOST	Over-sampling	۹۲/۰۲	۹۰/۰۵	۹۴/۰۶	۹۳/۰۸	۹۱/۰۲	۸۹/۰۷



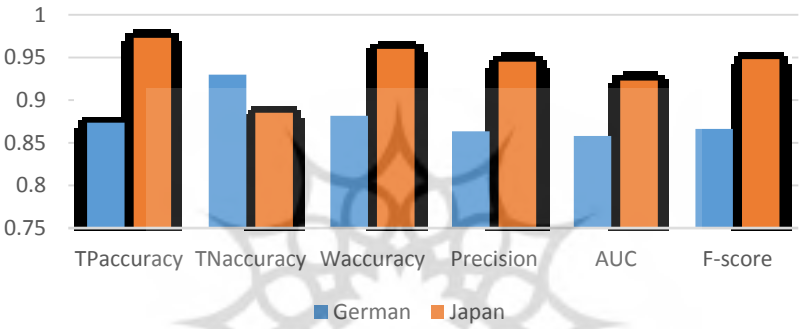
نمودار۸. Single SVM داده آلمان



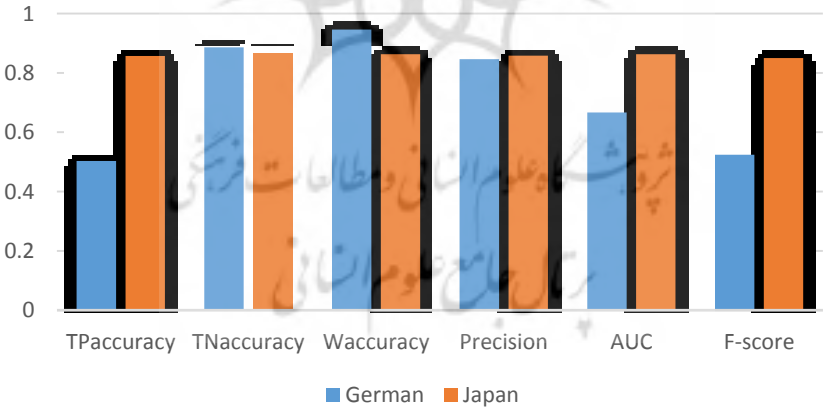
نمودار ۹. Single SVM داده ژاپن



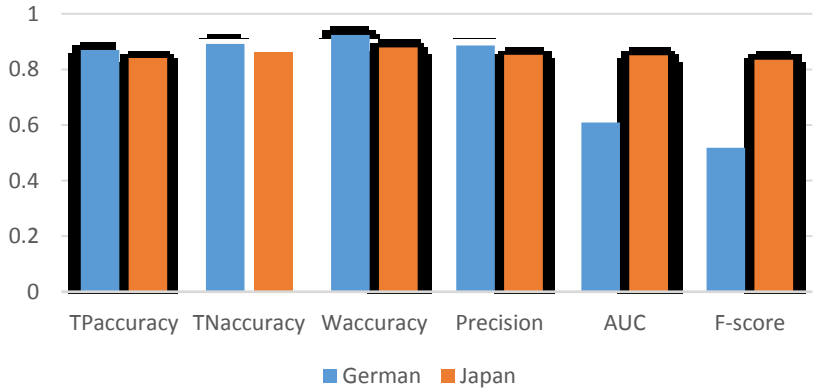
نمودار ۱۰. Majority voting آلمان و ژاپن



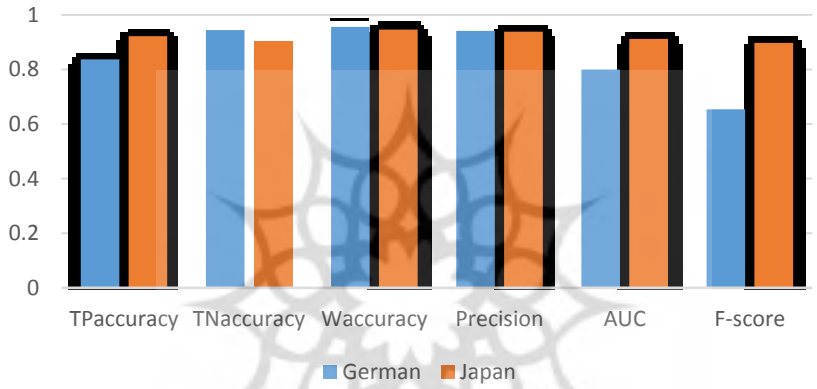
نمودار ۱۱. DBN-based آلمان و ژاپن



نمودار ۱۲. XGBoost آلمان و ژاپن



نمودار۱۳. Random Forest آلمان و ژاپن



نمودار۱٤. IDCOST آلمان و ژاپن

بحث و نتیجه‌گیری

نتایج این پژوهش نشان می‌دهد که مدل ترکیبی پیشنهادی، مبتنی بر ماشین بردار پشتیبان و شبکه باور عمیق، موفق به افزایش معنادار دقت طبقه‌بندی در داده‌های اعتباری نامتعادل شده است. یافته‌های کلیدی حاکی از آن است که رویکرد یکپارچه‌سازی انتخاب ویژگی با بهینه‌سازی پارامترهای کرنل RBF، به بهبود چشمگیری در عملکرد مدل منجر شده است. این بهبود در معیار میانگین دقت وزنی برای مجموعه‌داده آلمان حدود ۷/۵ درصد و برای مجموعه‌داده ژاپن حدود ۲ درصد در مقایسه با مدل‌های پایه مشاهده شد.

این نتایج، برتری روش پیشنهادی را در شرایط مشابه تأیید می‌کند. افزون بر این، به‌کارگیری روش انتخاب ویژگی به طبقه‌بندی و ارزش‌گذاری مؤثرتر متغیرهای ورودی کمک کرده و نقش مهمی در افزایش دقت نهایی مدل ایفا کرده است.

داده‌های اعتباری درنظر گرفته شده از مقالات مطالعه موردی استفاده شده است و محدودیت‌ها زمانیست که داده‌های واقعی اعتباری در اختیار ما قرارگرفته نشود.

مسیرهای پژوهشی آینده می‌تواند بر روی پیش‌بینی ریسک اعتباری، امتیازدهی داده‌های نامتعادل و همچنین طبقه‌بندی داده‌های خوش‌رفتار و بدرفتار باشد.

ملاحظات اخلاقی

پیروی از اصول اخلاق پژوهش

در این مقاله کلیه اصول اخلاقی و استانداردهای پژوهشی رعایت شده است.

مشارکت نویسندگان

تمام نویسندگان در طراحی، اجرا و نگارش این پژوهش مشارکت داشته‌اند.

تعارض منافع

نویسندگان هیچ‌گونه تعارض منافعی در خصوص این پژوهش ندارند.

حامی مالی

این پژوهش با هزینه شخصی انجام شده است و هیچ‌گونه کمک مالی از سازمان یا نهادی دریافت نکرده است.

References

Chang, Y. C; Chang, K. H; & Wu, G. J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, *73*, 914–920. https://doi.org/10.1016/j.asoc.2018.09.029

Danenas, P; & Garsva, G. (2012). Credit risk evaluation modeling using evolutionary linear SVM classifiers and sliding window approach. *Procedia Computer Science*, *9*, 1324–1333. https://doi.org/10.1016/j.procs.2012.04.145

Das, S; Meghanath, A; Behera, B. K; Mumtaz, S; Al-Kuwari, S; & Farouk, A. (2024). QFDNN: A resource-efficient variational quantum feature deep neural networks for fraud detection and loan prediction. *Expert Systems with Applications*, *258*, 122194. https://doi.org/10.1109/TCSS.2025.3568618

Ding, Y; Jia, M; Zhuang, J; & Ding, P. (2022). Deep imbalanced regression using cost-sensitive learning and deep feature transfer for bearing remaining useful life estimation. *Applied Soft Computing*, *127*, 109271. https://doi.org/10.1016/j.asoc.2022.109271

GhorbanniaDelavar, A; & Ziya, S. (2025). IDCOST: A Method for Increasing Data Criterion Service by Scoring Credit Imbalanced Data Using Applied SVM. *Journal of Modeling in Engineering*, *23*(81), 1–18. https://doi.org/10.22075/jme.2025.31252.2493. (In Persian).

GhorbanniaDelavar, A; Noori Lahrood, B; & Zekriyahpanah Gashti, M. (2011). ERPDRT: A novel real-time framework in integrated distributed system resources, secure with data mining mechanisms. *2011 IEEE 3rd International Conference on Communication Software and Networks*, 61–65. https://doi.org/10.1109/iccsn.2011.6013776

GhorbanniaDelavar, A; & Jafari, Z. (2015). A method for data classification reduction using weighting technique in SVM+. *Computer Science and Information Technology*, 13 (1), 162351e. https://jcsit.ir/article_162351.html. (In Persian).

Giusti, C; Guarnera, L; Casu, M; & Battiato, S. (2025). Fraud is not just rarity: A causal prototype attention approach to realistic synthetic oversampling. *Machine Learning with Applications*, *21*, 101345. https://doi.org/10.48550/arXiv.2507.14706

Gunnarsson, B. R; Vanden Broucke, S; Baesens, B; Óskarsdóttir, M; & Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, *295*(1), 292–305. https://doi.org/10.1016/j.ejor.2021.03.006

Gupta, R. K; Hassan, A; Majhi, S. K; Parveen, N; Zamani, A. T; Anitha, R; Ojha, B; Singh, A. K; & Muduli, D. (2025). Enhanced framework for credit card fraud detection using robust feature selection and a stacking ensemble model approach. *Results in Engineering*, *26*, 105084. https://doi.org/10.1016/j.rineng.2025.105084

Malek Mohammadi, M. R; Saeedi, A; & Matinfard, M. (2020). Investigating the systematic and unsystematic factors affecting credit risk in the Iranian banking system. *Quarterly Journal of Securities Exchange*, *13*(49), 134-159. https://doi.org/10.22034/jse.2020.11061.1317. (In Persian).

Mohagheghnia, M. J; Ghorbanizadeh, V; & Khanzadeh, M. (2021). Dimensions of credit rating of Iranian banks. *Quarterly Journal of Securities Exchange*, *14*(54), 63-84. https://doi.org/10.30495/jsed.2021.1921323.2038. (In Persian).

Sharma, R. (2025). AI-powered neural networks detecting anomalous patterns in real-time financial transactions. *Intelligent Systems with Applications*, *23*, 200037. https://doi.org/10.52783/jisem.v10i59s.12995

Tian, Z; Xiao, J; Feng, H; & Wei, Y. (2020). Credit Risk Assessment based on Gradient Boosting Decision Tree. *Procedia Computer Science*, *174*, 150–160. https://doi.org/10.1016/j.procs.2020.06.070

Yu, L; Yao, X; Wang, S; & Lai, K. (2011). Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection. *Expert Systems With Applications*, *38*(12), 15392–15399. https://doi.org/10.1016/j.eswa.2011.06.023

Yu, L; Zhou, R; Tang, L; & Chen, R. (2018). A DBN-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data. *Applied Soft Computing*, *69*, 192–202. https://doi.org/10.1016/j.asoc.2018.04.049

Zhang, Q; Wang, J; Lu, A; Wang, S; & Ma, J. (2017). An improved SMO algorithm for financial credit risk assessment – Evidence from China's banking. *Neurocomputing*, *272*, 314–325. https://doi.org/10.1016/j.neucom.2017.07.002

Zhang, T; Zhang, W; Xu, W; & Hao, H. (2018). Multiple instance learning for credit risk assessment with transaction data. *Knowledge-Based Systems*, *161*, 65–77.https://doi.org/10.1016/j.knosys.2018.07.030

Zhang, Y; Wang, G; Chung, F. L; & Wang, S. (2021). Support vector machines with the known feature-evolution priors. *Knowledge-Based Systems*, *223*, 107048. https://doi.org/10.1016/j.knosys.2021.107048