Original Article

Prediction of Ischemic Heart Disease with A Hybrid Algorithm of Support Vector Machine (SVM) and Xgboost Using Data from Yazd Health Study (YaHS)

Azam Barzegari [1], Mahyar Abdulmaleki ✉ [2], Masoud Mirzaei [3]

[1] Shahid Sadoughi University of Medical Sciences, Yazd, Iran

[2] M.Sc. in Computer Engineering – Software, Kurdistan, Iran

[3] Yazd Cardiovascular Research Centre, Non-Communicable Diseases Institute, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

✉ **Corresponding author:** M.Sc. in Computer Engineering – Software, Kurdistan, Iran

Email: abdolmalakmahyar@gmail.com
Tel: +989187870758
Postal Code: 6661116915

Received: 8 Nov 2026
Revised: 5 May 2026
Accepted: 31 May 2026

Citation: Barzegari A, Abdulmaleki M, Mirzaei M. Prediction of Ischemic Heart Disease with a Hybrid Algorithm of Support Vector Machine (SVM) and XGBoost Using Data from Yazd Health Study (YaHS). J Shahid Sadoughi Uni Med Sci 2026; 34(7): 10451-463.

This work is licensed under a Creative Commons Attribution 4.0 Unported License which allows users to read, copy, distribute and make derivative works for non-commercial purposes from the material, as long as the author of the original work is cited properly.

Abstract

Introduction: Ischemic heart disease represents one of the most prevalent conditions characterized by elevated mortality rates. The aim of this research was to provide an efficient instrument for the early diagnosis of ischemic heart disease and to assist clinical decision-making processes through the application of hybrid artificial intelligence algorithms.

Methods: The data used in this study were from Yazd Health Study (YaHS) enrollment phase dataset, comprising 9969 participants and 30 selective features from a total of 300, collected between 2014 to 2015. Data were analyzed using SVM and XGBoost machine learning algorithms using Python version 3.11.8.

Results: Combining the strengths of SVM and XGBoost models through a Voting Classifier approach enhanced both the accuracy and efficiency of the heart disease prediction model. When evaluating the predictive performance for heart disease, the Voting Classifier model achieved in accuracy rate of 90%.

Conclusion: These findings suggest that implementing machine learning techniques and hybrid algorithms for early diagnosis and prediction could facilitate earlier intervention in CHD, ultimately helping to lower mortality rates.

Keywords: Prediction, Ischemic heart disease, SVM, XGBoost, Voting Classifier, Yazd Health Study.

[Downloaded from jssu.ssu.ac.ir on 2026-10-21]

[DOI: 10.18502/ssu.v34i7.22992]

Extended Abstract

Introduction

Ischemic heart disease (IHD) remains one of the most prevalent cardiovascular conditions and a leading cause of mortality worldwide. Early and accurate diagnosis is critical because timely intervention can significantly reduce severe complications, mortality rates, and improve patients' quality of life. Traditional diagnostic methods often require invasive procedures and impose considerable time and financial burdens on healthcare systems. In recent years, remarkable advances in machine learning and the availability of largescale health data have created new opportunities for developing intelligent prediction models. These models can analyze complex patterns within data to identify atrisk individuals and assist physicians in making more informed clinical decisions. The aim of this research was to provide an efficient instrument for the early diagnosis of ischemic heart disease and to assist clinical decisionmaking processes through the application of hybrid artificial intelligence algorithms. This study specifically sought to combine the strengths of Support Vector Machine (SVM) and XGBoost models using a Voting Classifier approach to enhance both the accuracy and efficiency of heart disease prediction.

Methods

The data used in this study were derived from the Yazd Health Study (YaHS) enrollment phase dataset, comprising 9,969 participants and 30 selective features from a total of 300, collected between 2014 and 2015. The YaHS is a populationbased cohort study of adults aged 2070 years in Yazd, Iran. Data were analyzed using SVM and XGBoost machine learning algorithms implemented in Python version 3.11.8. The dataset was divided into three parts: 70% for training, 15% for validation, and 15% for testing. The training set was used for learning model parameters, the validation set for hyperparameter tuning and model selection, and the test set for final performance evaluation on unseen data. Data splitting was performed randomly while preserving class ratios to prevent bias and data leakage.

Preprocessing steps included handling missing values through imputation methods such as mean, median, or mode replacement, and standardization using Zscore normalization to ensure that all features contributed equally to the model. Feature selection was performed using filter methods, wrapper methods, and embedded methods to identify the most relevant predictors and reduce model complexity.

For hyperparameter tuning, both SVM and XGBoost models were optimized using Grid Search with KFold CrossValidation on the training and validation sets. The SVM parameters including C and gamma, and XGBoost parameters including learning rate, n_estimators, colsample_bytree, subsample, and max_depth were systematically evaluated. The bestperforming hyperparameters were selected based on F1Score and ROC AUC metrics.

The hybrid Voting Classifier model was constructed by combining the optimized SVM and XGBoost models as base models. Soft Voting was employed, in which each model predicts the probability of sample assignment to each class, and the average of these probabilities is calculated for each class. The class with the highest average probability is selected as the final prediction. This approach typically performs better than Hard Voting because it considers more information (probabilities).

Evaluation metrics included Accuracy, Precision, Recall, F1Score, and ROC AUC. Statistical analysis was conducted using Python version 3.11.8, along with Scikitlearn version 1.4.1, NumPy version 1.26.4, Pandas version 2.2.1, and XGBoost version 2.0.3.

Results

The results demonstrated that combining the strengths of SVM and XGBoost models through a Voting Classifier approach enhanced both the accuracy and efficiency of the heart disease prediction model. The evaluation results for each model were as follows: The XGBoost model achieved an accuracy of 0.88, precision of 0.87, recall of 0.86, F1Score of 0.865, and ROC AUC of 0.91. The SVM model achieved an accuracy of 0.85, precision of 0.83, recall of 0.84, F1Score of 0.835, and ROC AUC of 0.89. The hybrid Voting Classifier model outperformed both individual models, achieving an accuracy of 0.90, precision of 0.89, recall of 0.88, F1Score of 0.885, and ROC AUC of 0.92.

Feature importance analysis using XGBoost revealed that the most important features for predicting ischemic heart disease, in order of

[Downloaded from jssu.ssu.ac.ir on 2026-10-21]

[DOI: 10.18502/ssu.v34i7.22992]

CC BY-NC 4.0

importance, were: diabetes, family history of disease, high cholesterol, high blood pressure, age, diastolic blood pressure (DBP), systolic blood pressure (SBP), and body mass index (BMI). Based on classification, eight superior features were selected for the modeling stage. Ultimately, age was identified as one of the most important and influential factors in predicting ischemic heart disease across all models.

The comparison of model performance metrics across SVM, XGBoost, and Voting Classifier models demonstrated that the Voting Classifier consistently achieved the highest values across all metrics. ROC AUC analysis showed that the Voting Classifier achieved the highest area under the curve (0.92), followed by XGBoost (0.91) and SVM (0.89). The ROC curve analysis confirmed that the Voting Classifier model had the highest diagnostic power and was most capable of distinguishing between patients and healthy individuals, indicating that this model made the fewest diagnostic errors.

The findings of this study are consistent with previous research demonstrating the effectiveness of machine learning algorithms in predicting heart disease. Studies by Mahmoodi (2017) using SVM achieved 85% accuracy, Setiyadi et al. (2025) using SVM achieved 85% accuracy, and Gandha and Richhariya (2024) using XGBoost achieved 98.04% accuracy. The present study's hybrid approach demonstrated competitive performance while leveraging a larger, populationbased dataset from the Yazd Health Study.

Compared to the first part of this study's results, which used a combination of Random Forest, Naïve Bayes, and 20 parameters with RapidMiner software and achieved 74.51% accuracy for heart disease prediction, the present study demonstrated significantly better and higher accuracy. This improvement can be attributed to the advanced hybrid algorithm, precise preprocessing, and comprehensive hyperparameter tuning.

Conclusion

These findings suggest that implementing machine learning techniques and hybrid algorithms for early diagnosis and prediction could facilitate earlier intervention in coronary heart disease, ultimately helping to lower mortality rates. The Voting Classifier model, which combines SVM and XGBoost through Soft Voting, demonstrated superior performance compared to individual models across all evaluation metrics, including Accuracy, Precision, Recall, F1Score, and ROC AUC. Achieving high evaluation metric values (ROC AUC = 0.92) reveals the considerable potential of this model for use in early diagnosis of ischemic heart disease. Early and accurate diagnosis of ischemic heart disease is of high importance, and the developed model can serve as an efficient tool for physicians and health professionals to identify atrisk individuals, enabling timely intervention, reducing mortality, and improving patients' quality of life. Despite limitations such as interpretability and the need for validation on more diverse data, this research represents an important step toward using innovative machine learning technologies to improve community health, particularly in the field of prediction and prevention of ischemic heart diseases. Future research can further develop this field by examining more advanced ensemble algorithms, deeper interpretability methods, and using broader datasets.

[Downloaded from jssu.ssu.ac.ir on 2026-10-21]

[DOI: 10.18502/ssu.v34i7.22992]

CC BY-NC 4.0

مقاله پژوهشی

پیش‌بینی بیماری‌های ایسکمیک قلبی با الگوریتم ترکیبی Support Vector Machine (SVM) و XGBoost با استفاده از داده‌های مطالعه سلامت مردم یزد (یاس)

اعظم برزگری[۱]، مهیار عبدالملکی[۲]✉، مسعود میرزایی[۳]

[۱] معاونت تحقیقات و فناوری- دانشگاه علوم پزشکی شهید صدوقی، یزد، ایران
[۲] کارشناسی ارشد مهندسی کامپیوتر- نرم افزار، کردستان ایران
[۳] مرکز تحقیقات قلب و عروق، پژوهشکده بیماری‌های غیرواگیر دانشگاه علوم پزشکی شهید صدوقی، یزد، ایران

چکیده

مقدمه: یکی از شایع‌ترین بیماری‌ها با مرگ و میر بالا بیماری های ایسکمیک قلبی می‌باشد. هدف از این پژوهش ارائه ابزاری کارآمد برای تشخیص زودهنگام بیماری ایسکمیک قلبی و کمک به تصمیم‌گیری‌های بالینی، با استفاده از الگوریتم‌های ترکیبی هوش مصنوعی می‌باشد.

روش بررسی: داده‌های مورد استفاده در این تحقیق از فاز ثبت نام مطالعه سلامت مردم یزد (یاس) که شامل ۹۹۶۹ شرکت کننده و ۳۰ ویژگی از ۳۰۰ ویژگی این مجموعه داده جمع آوری شده از سال ۱۳۹۳ تا ۱۳۹۴ بود. داده ها با استفاده از الگوریتم‌های یادگیری ماشین SVM و XGBoost توسط نرم‌افزار پایتون نسخه ۳.۱۱.۸ تجزیه و تحلیل گردید.

نتایج: با ترکیب نقاط قوت مدل‌های SVM و XGBoost و با به کارگیری رویکرد Voting Classifier دقت و کارایی مدل پیش‌بینی بیماری ایسکمیک قلب بهبود یافت. نتایج تجزیه و تحلیل داده‌ها با استفاده از مدل Voting Classifier در پیش‌بینی بیماری ایسکمیک قلبی، دقت ۹۰٪ را نشان داد.

نتیجه‌گیری: با استفاده از تکنیک یادگیری ماشین و الگوریتم‌های ترکیبی و با تشخیص و پیش‌بینی زودهنگام بیماری، می توان درمان مناسب را در مراحل اولیه انجام داد و باعث کاهش مرگ‌ومیر بیماران شد.

واژه‌های کلیدی: ایسکمیک قلبی، SVM، XGBoost، مطالعات سلامت مردم یزد

✉ **نویسنده مسئول:**
مهیار عبدالملکی
کارشناسی ارشد مهندسی کامپیوتر- نرم افزار، کردستان ایران

پست الکترونیک:
abdolmalakmahyar@gmail.com
تلفن: ۰۹۱۸۷۸۷۰۷۵۸
صندوق پستی: ۶۶۶۱۱۱۶۹۱۵

تاریخ دریافت: ۱۴۰۵/۸/۱۷
تاریخ بازنگری: ۱۴۰۵/۲/۱۵
تاریخ پذیرش: ۱۴۰۵/۳/۱۰

ارجاع:
برزگری اعظم، عبدالملکی مهیار، میرزایی مسعود. پیش‌بینی بیماری‌های ایسکمیک قلبی با الگوریتم ترکیبی Support Vector Machine (SVM) و XGBoost با استفاده از داده مطالعه سلامت مردم یزد (یاس). مجله علمی پژوهشی دانشگاه علوم پزشکی شهید صدوقی یزد ۱۴۰۵؛ ۳۴ (۷): ۴۶۳-۱۰۴۵۱

[Downloaded from jssu.ssu.ac.ir on 2026-10-21]

[DOI: 10.18502/ssu.v34i7.22992]

CC BY-NC 4.0

مقدمه

بیماری‌ها قلبی عروقی یکی از علل اصلی مرگ‌ومیر در سراسر جهان محسوب می‌شوند. تشخیص زودهنگام و دقیق این بیماری‌ها از اهمیت بالایی برخوردار است، زیرا مداخله به‌موقع باعث کاهش عوارض شدید- مرگ و میر و بهبود کیفیت زندگی بیماران می‌شود (۱). بیماری قلبی را می‌توان به هفت نوع اصلی تقسیم کرد: آریتمی، بیماری مادرزادی قلب، کاردیومیوپاتی، میوکاردیت، بیماری عروق کرونر، آنژین صدری، انفارکتوس میوکارد و نارسایی قلب که سه نوع اخیر، بیماری‌های ایسکمیک قلب شناخته می‌شوند. خطر ابتلا به این بیماری در سطوح مختلف ارزیابی شد و نشان داد دوره‌های درمانی در سطح خطر اولیه، هزینه‌های کمتری داشته و باعث افزایش طول عمر بیماران می‌شود (۲). روش‌های سنتی تشخیص اغلب نیازمند معاینات تهاجمی بوده و زمان و هزینه قابل‌توجهی را به همراه دارند. اما در سال‌های اخیر، پیشرفت‌های چشمگیر در حوزه یادگیری ماشین و دسترسی به حجم وسیعی از داده‌های سلامت، فرصت‌های جدیدی را برای توسعه مدل‌های پیش‌بینی هوشمند فراهم کرد. این مدل‌ها با تحلیل الگوهای پیچیده موجود در داده‌ها، قادرند افراد در معرض خطر را شناسایی کرده و به پزشکان در اتخاذ تصمیمات آگاهانه‌تر یاری رسانند. در تحقیق فصیحی و همکاران از الگوریتم ماشین بردار پشتیبان برای تشخیص بیماری عروق کرونر استفاده شد. تجزیه و تحلیل داده‌ها با استفاده از نرم‌افزار MATLAB و در سطح معنی‌داری ۰/۰۵ انجام گرفت. یافته‌ها نشان داد که با بهره‌گیری از سوابق پزشکی شامل ۱۴ ویژگی مشترک، مرتبط با اطلاعات آنتروپومتری، تست‌های تشخیصی، نتایج آنژیوگرافی و فعالیت بدنی، الگوریتم ماشین‌بردار پشتیبان بیماری عروق کرونر را با دقت ۷۰٪ پیش‌بینی کرد (۳). در پژوهش سیاه‌کوهیان و همکاران از الگوریتم آنالیز افتراقی برای تشخیص بیماری عروق کرونر استفاده شد. نرم‌افزار پایتون نیز برای تجزیه و تحلیل داده‌ها به‌کار گرفته شد. یافته‌ها نشان داد که با استفاده از ۱۴ ویژگی به‌عنوان عوامل خطر مربوط به اطلاعات آزمایشگاهی، شخصی و سبک زندگی آزمودنی‌ها، الگوریتم آنالیز افتراقی بیماری عروق کرونر را با دقت ۹۴/۴٪ پیش‌بینی می کند (۴). طباطبایی و همکاران با استفاده از الگوریتم‌های شبکه عصبی، مبتی بر قانون، رگرسیون و آدابوست جهت ارزیابی مجموعه داده‌های یاس با ۱۴ پارامتر انجام دادند که نتایج تجزیه و تحلیل نشان داد که الگوریتم CN2 با دقت ۸۶/۳٪ بهترین عملکرد را داشته است (۵).

استفاده از الگوریتم‌های یادگیری ماشین با تنظیم پارامترهای پیشرفته در دو حالت بررسی شد: حالت اول بدون ویژگی‌های جغرافیایی و حالت دوم با ویژگی‌های جغرافیایی که شامل رطوبت هوا، دما و وضعیت آموزش یک مکان بودند. عملکرد دو آزمایش با استفاده از معیارهای طبقه‌بندی مقایسه شد. یافته‌ها نشان داد که عملکرد آزمایش حالت دوم بهتر از آزمایش اول بود. الگوریتم XGBoost به بالاترین دقت ۹۵/۲۴٪ دست یافت (۶). در مطالعات قبلی الگوریتم طبقه‌بندی و درخت رگرسیون، یک روش یادگیری ماشینی نظارت‌شده، برای پیش‌بینی بیماری قلبی و استخراج قوانین تصمیم‌گیری به منظور شفاف‌سازی روابط بین متغیرهای ورودی و خروجی استفاده شده است. علاوه بر این، یافته‌های این مطالعه ویژگی‌های موثر بر بیماری قلبی را بر اساس اهمیت رتبه‌بندی می‌کند. عملکرد، دقت ۸۷٪ پیش‌بینی اعتبار مدل را تأیید می‌کند (۷). پژوهش پیشین با استفاده از الگوریتم‌های کلاسیک نظیر درخت تصمیم و نایبو بیز (Naïve Bayes) در محیط‌های گرافیکی مانند RapidMiner متمرکز بوده‌اند. این الگوریتم‌ها اگرچه از سادگی، تفسیرپذیری، و سرعت اجرای بالایی برخوردارند، اما در تحلیل داده‌های پیچیده، پرتراکم و دارای تعاملات چندمتغیره دچار محدودیت هستند. از سوی دیگر، محیط‌های گرافیکی معمولاً انعطاف کافی برای تنظیم دقیق پارامترها، کنترل فرآیند یادگیری، یا بازتولید کامل نتایج را ندارند. در نتیجه، خروجی این نوع مطالعات بیشتر در سطح تحلیل توصیفی باقی مانده و توانایی آن‌ها در پیش‌بینی دقیق پدیده‌های پیچیده محدود است. در مقابل، پژوهش حاضر با بهره‌گیری از ترکیب الگوریتم‌های XGBoost و SVM در محیط Python، از یک چارچوب تحلیلی پیشرفته و داده‌محور استفاده کرده که چندین تفاوت بنیادین و شکاف دانشی نسبت به رویکردهای پیشین ایجاد می‌کند: افزایش دقت و کارایی مدل؛ ارتقای قابلیت تعمیم‌پذیری، بهبود بازتولیدپذیری، تحلیل عمیق اهمیت ویژگی‌ها و حرکت از تحلیل توصیفی به پیش‌بینی‌کننده این پژوهش در عبور از روش‌های سنتی مبتنی بر ابزارهای گرافیکی به سمت مدل‌های ترکیبی، خودتنظیم و مبتنی بر کد در محیط‌های علمی باز (Open Source) است. این گذار نه‌تنها موجب ارتقای دقت و پویایی تحلیل‌ها می‌شود، بلکه بنیانی برای توسعه الگوهای بومی هوش مصنوعی قابل تفسیر در حوزه

CC BY-NC 4.0

[Downloaded from jssu.ssu.ac.ir on 2026-10-21]
[DOI: 10.18502/ssu.v34i7.22992]

مطالعات داده‌محور فراهم می‌آورد. هدف اصلی این تحقیق، توسعه و ارزیابی مدل ترکیبی یادگیری ماشین بر پایه الگوریتم‌های SVM و XGBoost با استفاده از داده‌های مطالعه سلامت مردم یزد (یاس) به منظور پیش‌بینی دقیق‌تر بیماری‌های ایسکمیک قلبی و شناسایی زودهنگام بیماران مستعد در کوتاه‌ترین زمان است.

روش بررسی

پژوهش حاضر یک مطالعه کاربردی است که به پیش‌بینی بیماری قلبی پرداخته است. جامعه آماری مورد استفاده در تحقیق حاضر از مجموعه داده‌های فاز ثبت نام مطالعه سلامت مردم یزد (یاس) می‌باشد. نمونه‌گیری مطالعه یاس به‌صورت خوشه‌ای در سطح شهر یزد انجام گرفته و درمرحله اول نمونه‌گیری دسته‌بندی برحسب کدپستی و خوشه‌های شهری صورت گرفت که ۲۰۰ خوشه ۵۰ نفری انتخاب شدند. مطالعه یاس، یک مطالعه کوهورت آینده‌نگر مبتنی بر جمعیت است که در شهر یزد برروی ۹۹۶۹ نفر بین سنین ۶۹-۲۰ سال از سال ۱۳۹۳ در حال انجام است. فاز اول آن در طول سال‌های ۱۳۹۳ تا ۱۳۹۴ انجام شده و فاز دوم آن در سال ۱۴۰۴ تکمیل شد و فالوآپ آن همچنان در دست اجرا است. مرحله اول نمونه‌گیری از اطلاعات خانوارها بصورت پرسشنامه‌ای بود و اطلاعات مورد نیاز به‌صورت مصاحبه و تکمیل پرسش‌نامه انجام شده و پرسش‌نامه‌ها دارای پاسخ نامه الکترونیکی قابل خوانده شدن توسط رایانه بود. فشارخون، قد و وزن افراد در منازل اندازه‌گیری شد و به افراد دعوتنامه داده شد تا در آزمایشگاه مرکزی جهت تحویل نمونه خون حضور یافته تا اطلاعات کامل‌تری جمع‌آوری شود. اعتبار صوری پرسش‌نامه‌ها مورد بررسی قرار گرفته و روی ۵۰ شرکت‌کننده، پایلوت شد، الفای کرونباخ برابر ۸۱٪ بود بنابراین پرسشنامه معتبر در نظر گرفته شد (۸). مجموعه داده‌ها شامل ۳۰ ویژگی (سن، جنس، میزان تحصیلات، وضعیت شغلی و غیره) بود که از مجموعه داده‌های دیتاست فاز ثبت نام مطالعه یاس اخذ شد یکی از این فاکتورها، فاکتور خطر بیماری‌های ایسکمیک قلبی می‌باشد. این مطالعه از الگوریتم ترکیبی SVM و XGBoost با رویکرد Classifier-Voting، استفاده کرده است. ماشین بردار پشتیبان (Support Vector Machine) یک الگوریتم یادگیری ماشین قدرتمند و پرکاربرد برای مسائل طبقه‌بندی و رگرسیون است. در مسائل طبقه‌بندی، هدف SVM یافتن یک ابرصفحه (Hyperplane) است که داده‌ها را با حداکثر حاشیه (Margin) بین کلاس‌های مختلف تفکیک کرده و در مسائل خطی و غیرخطی قابل استفاده کند. در فضاهای خطی قابل تفکیک، این ابرصفحه منحصربه‌فرد است. اما در بسیاری از مسائل واقعی، داده‌ها به‌صورت خطی قابل تفکیک نیستند.

معادله اصلی برای تصمیم‌گیری در SVM (برای دو کلاس) به صورت زیر است:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right)$$

که در آن x نمونه ورودی، x_i بردارهای پشتیبان، y_i برچسب کلاس بردارهای پشتیبان، K تابع کرنل، α_i ضرایب بهینه‌سازی شده، و b بایاس است. مزایای الگوریتم SVM شامل مواردی از جمله پایه ریاضی معتبر؛ توانایی طبقه‌بندی داده‌های خطی و غیرخطی؛ قابلیت استفاده از توابع کرنل؛ قابلیت ترکیب با مدل‌های دیگر؛ کارایی بالا در داده‌های با ابعاد زیاد می‌باشد (۹).

XGBoost (Extreme Gradient Boosting) یک الگوریتم قدرتمند مبتنی بر درخت تصمیم است که با استفاده از تکنیک «تقویت گرادیانی» مدل‌هایی دقیق و سریع تولید می‌کند که به‌دلیل سرعت بالا، عملکرد عالی و توانایی مدیریت داده‌های بزرگ شهرت دارد Gradient Boosting یک روش Ensemble است که مدل‌های ضعیف (معمولاً درختان تصمیم) را به‌صورت متوالی آموزش می‌دهد. هرمدل جدید، سعی در تصحیح خطاهای مدل قبلی دارد. در Gradient Boosting، مدل جدید با پیش‌بینی باقیمانده (Residual) مدل قبلی آموزش داده می‌شود.

تابع هدف (Loss Function) در XGBoost معمولاً به‌صورت زیر است:

$$\text{Objective} = \sum_{i=1}^{N} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

که در آن L تابع زیان(مانند زیان Logloss برای طبقه‌بندی)، $y^\wedge_i$ پیش‌بینی مدل، y_i مقدار واقعی، f_k درخت k-ام، و

$$(f_k) = \gamma\|f_k\|_2 + \tfrac{1}{2}\lambda\|f_k\|_2^2$$

Ω تابع Regularization است که پیچیدگی درخت را کنترل می‌کند. γ و λ پارامترهای Regularization هستند.

CC BY-NC 4.0
[Downloaded from jssu.ssu.ac.ir on 2026-10-21]
[DOI: 10.18502/ssu.v34i7.22992]

مزایای الگوریتم XGBoost شامل مواردی از جمله دقت و سرعت بالا-جلوگیری از بیش‌برازش-پشتیبانی از داده‌های ناقص-انعطاف‌پذیری بالا می‌باشد (۱۰).

Voting Classifier یک استراتژی Ensemble است که نتایج پیش‌بینی چندین مدل پایه (Besa Models) را با هم ترکیب می‌کند تا به یک پیش‌بینی نهایی برسد. دو رویکرد اصلی برای ترکیب پیش‌بینی‌ها شامل Soft Voting و Hard Voting (Majority Voting) در Voting Classiifer وجود دارد:

Soft Voting: در این روش، هرمدل احتمال انتساب نمونه به هرکلاس را پیش‌بینی می‌کند. سپس، میانگین این احتمالات برای هرکلاس محاسبه می‌شود و کلاسی که بالاترین میانگین احتمال را داشته باشد، به‌عنوان پیش‌بینی نهایی انتخاب می‌شود. این روش معمولاً عملکرد بهتری نسبت به Hard Voting دارد، زیرا اطلاعات بیشتری(احتمالات) را درنظر می‌گیرد.

$$P(\text{class } c) = \frac{1}{K}\sum_{k=1}^{K} P_k(\text{class } c)$$

$$\hat{y} = \arg\max_{c} P(\text{class } c)$$

که در آن $P_k(\text{class } c)$ احتمال انتساب به کلاس c توسط مدل-kام است.

جدول۱: ماتریس درهم ریختگی

واقعی	پیش‌بینی Positive	Negative	مجموع
Positive	TP	FN	P=TP+FN
Negative	FP	TN	N=FP+TN
Total	TP+FP	FN+TN	

جدول۲: نحوه محاسبه معیارهای ارزیابی

Evaluation Methods	Equations
Accuracy	$\frac{(TP+TN)}{(TP+FN+FP+TN)}$
Precision	$\frac{TP}{(TP+FP)}$
Recall	$\frac{TP}{(TP+FN)}$
F Score	$\frac{2\times(Precision\times Recall)}{(Precision+Recall)}$

معیارهای ارزیابی: در این پژوهش برای ارزیابی معیارهای دقت کلی(Accuracy)؛ Precision (دقت مثبت)؛ Recall (حساسیت یا پوشش) و F1-Score از ماتریس درهم ریختگی (Confusion Matrix) استفاده شد.

جهت ارزیابی دسته‌ها از مقادیر ماتریس درهم ریختگی استفاده می‌شود (جدول۱). نحوه محاسبه معیارهای ارزیابی را براساس ماتریس درهم ریختگی نشان می‌دهد (جدول ۲).

این مطالعه از نوع CRISP: Cross-industry standard process می‌باشد که از معروف‌ترین و متداول‌ترین روش برای انجام پروژه‌های هوش مصنوعی، به‌خصوص در حوزه پیش‌بینی بیماری‌ها است. هم‌چنین دلیل استفاده از این متد علاوه بر قابل‌فهم بودن برای کاربران مبتدی، به استناد تحقیقات انجام شده توسط انجمن تحقیقات داده‌کاوی در سال ۲۰۰۷ پرکاربردتر بودن از سایر روش می‌باشد (۱۱). قبل از اعمال الگوریتم‌های

CC BY-NC 4.0
[Downloaded from jssu.ssu.ac.ir on 2026-10-21]
[DOI: 10.18502/ssu.v34i7.22992]

یادگیری ماشین، انجام مراحل پیش‌پردازش بر روی داده‌ها امری ضروری است. مراحل پیش‌پردازش شامل موارد زیر است:

مدیریت مقادیر گمشده (Missing Value Imputation): برای مدیریت این مقادیر، از رویکردهای مختلفی از جمله حذف، جایگزینی با میانگین/میانه، جایگزینی با نما؛ و روش‌های پیشرفته‌تر استفاده شد.

استاندارد سازی (Standardization): بسیاری از الگوریتم‌های یادگیری ماشین، به‌ویژه الگوریتم‌های مبتنی بر فاصله یا گرادیان، به‌مقیاس ویژگی‌ها حساس هستند. اگر مقادیر ویژگی‌ها در محدوده‌های بسیار متفاوتی قرار داشته باشند، ویژگی‌هایی با مقادیر بزرگتر ممکن است بر خروجی مدل تأثیر نامتناسبی داشته باشند. استانداردسازی، مقادیر ویژگی‌ها را به گونه ای تغییر می‌دهد که میانگین آنها صفر و واریانس آن‌ها یک شود (Z-score Standardization):

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}$$

که در آن x مقدار اصلی، μ میانگین ویژگی و σ انحراف معیار آن ویژگی است. این فرآیند تضمین می کند که تمام ویژگی‌ها در مقیاس مشابهی قرار گیرند.

انتخاب ویژگی (Feature Selection): در صورت وجود تعداد بسیار زیادی ویژگی، انتخاب مهم‌ترین آنها می‌تواند به کاهش پیچیدگی مدل، جلوگیری از بیش‌برازش و بهبود زمان آموزش کمک کند. روش‌های مختلفی برای انتخاب ویژگی وجود دارد، از جمله: روش‌های فیلتر(Filter Methods) شامل روش‌های پوششی (Wrapper Methods) و روش‌های تعبیه شده (Embedded Methods) می‌باشد.

برای ارزیابی الگوریتم‌ها راه‌های مختلفی وجود دارد از جمله: روش آزمایشگاهی، روش ریاضی و روش پیاده‌سازی و با توجه به معیارهای ارزیابی که وجود دارد می‌توان بهترین و مناسب‌ترین روش را انتخاب نمود زیرا ارزیابی الگوریتم‌ها در دنیای واقعی کار دشواری است و اینکه الگوریتم پیشنهادی جزو الگوریتم مسائل طبقه‌بندی و پیش‌بینی باشد و معیارهای دقت و صحت مهم‌ترین معیارها برای پیش‌بینی می‌باشد از ماتریس سردرگمی می‌توان استفاده کرد (۱۲).

منحنی ROC AUC (Area Under the Receiver Operating Characteristic Curve): ROC یک نمودار است که نرخ مثبت واقعی (True Positive Recall یا Rate) را در برابر نرخ مثبت کاذب(False Positive Rate) یا $\frac{FP}{FP+TN}$ در آستانه‌های مختلف طبقه‌بندی رسم می‌کند. مساحت زیر این منحنی (AUC) معیاری برای سنجش توانایی کلی مدل در تمایز بین کلاس‌های مثبت و منفی است. مقدار AUC بین ۰.۵ (عملکرد تصادفی) تا ۱ (عملکرد کامل) متغیر است AUC بالاتر نشان‌دهنده توانایی بهتر مدل در تفکیک کلاس‌ها است (۱۳).

معماری مدل و فرآیند آموزش شامل این مراحل است تقسیم داده‌ها: مجموعه داده سلامت مردم یزد (یاس) به سه بخش تقسیم شد مجموعه آموزش، اعتبارسنجی و مجموعه تست یا آزمون. معمولاً نسبت‌های ۷۰-۱۵-۱۵ یا ۸۰-۱۰-۱۰ استفاده می‌شود. مجموعه آموزش برای یادگیری پارامترهای مدل‌ها، مجموعه اعتبارسنجی برای تنظیم هایپرپارامترها و انتخاب بهترین مدل و مجموعه آزمون برای ارزیابی نهایی عملکرد مدل در داده‌های دیده نشده به‌کار رفت.

تنظیم هایپرپارامترها: برای هر یک از مدل‌های SVM و XGBoost هایپرپارامترهای کلیدی با استفاده از جستجوی شبکه‌ای(Grid Search) همراه با اعتبارسنجی متقابل (K-Fold Cross-Validation) بر روی مجموعه آموزش و اعتبارسنجی تنظیم شدند. به عنوان مثال، برای SVM پارامترهای C و gamma و برای XGBoost پارامترهایی مانند n_estimators, learning_rate, colsample_bytree, max_depth, subsample، بررسی شدند. هدف یافتن مجموعه‌ای از هایپرپارامترها بود که بهترین عملکرد را در مجموعه اعتبارسنجی (براساس معیاری مانند F1-Score یا ROC AUC) نشان دهد.

آموزش مدل‌های پایه: پس از تعیین بهترین هایپرپارامترها، مدل‌های SVM و XGBoost با استفاده از این پارامترها برروی کل مجموعه آموزش(آموزش + اعتبارسنجی) دوباره آموزش داده شدند. ساخت VotingClassiifer دو مدل بهینه شده SVM و XGBoost به‌عنوان مدل‌های پایه انتخاب شده و در VotingClassiifer با روش (Soft Voting) ترکیب شدند.

ارزیابی نهایی: عملکرد مدل‌های منفرد SVM، XGBoost و مدل ترکیبی Voting Classiifer بر روی مجموعه آزمون(Test Set) که در هیچ مرحله‌ای از آموزش یا تنظیم هایپرپارامتر استفاده نشده بود، با استفاده از معیارهای ذکر شده ارزیابی شد.

CC BY-NC 4.0
[Downloaded from jssu.ssu.ac.ir on 2026-10-21]
[DOI: 10.18502/ssu.v34i7.22992]

تحلیل نتایج: نتایج حاصل از معیارهای ارزیابی برای هر سه مدل با یکدیگر مقایسه شدند تا اثربخشی رویکرد ترکیبی در پیش‌بینی بیماری قلبی سنجیده شود.

تجزیه و تحلیل آماری

از برنامه پایتون نسخه 3.11.8 جهت انجام این تحقیق استفاده شد. استفاده از این نسخه‌ها اطمینان می‌دهد که از آخرین ویژگی‌های زبان ماشین و بهبودهای پایدار استفاده شده است (۱۴). هم‌چنین از نسخه 1.4.1 کتابخانه Scikit-learn، نسخه 2.2.1 کتابخانه Pandas و نسخه 1.26.4 کتابخانه Numpuy استفاده شده است. نسخه مورد استفاده XGBoost 2.0.3 است.

نتایج

در این پژوهش، داده‌ها به سه بخش آموزش، اعتبارسنجی و آزمون تقسیم شدند. نسبت ۷۰ درصد برای آموزش، ۱۵ درصد برای اعتبارسنجی و ۱۵ درصد برای آزمون درنظر گرفته شد. مجموعه آموزش برای یادگیری پارامترهای مدل، مجموعه اعتبارسنجی برای تنظیم ابرپارامترها و انتخاب بهترین مدل، و مجموعه آزمون برای ارزیابی نهایی عملکرد مدل در داده‌های دیده‌نشده مورد استفاده قرار گرفت. تقسیم داده‌ها به‌صورت تصادفی و با حفظ نسبت کلاس‌ها انجام شد تا از سوگیری و نشت اطلاعات جلوگیری شود. مجموعه داده‌های استفاده شده در این تحقیق از مجموعه داده‌های فاز ثبت نام مطالعه سلامت مردم یزد (یاس) می‌باشد که با استفاده از الگوریتم‌های SVM و XGBoost برای پیش‌بینی بیماری‌های ایسکمیک قلبی استفاده شد. برای بهبود و نوآوری یک مدل ترکیبی از این دو الگوریتم را اجرا کردیم که نتایج نشان داد تشخیص بیماری ایسکمیک قلبی با استفاده از الگوریتم ترکیبی بهتر خواهد بود. مقدار دقت مدل‌های تولید شده برای داده‌های آموزشی و آزمایش در (جدول ۳) نشان داده شده است که SVM (۸۵٪)، XGBoost (۸۸٪) و مدل ترکیبی (۹۰٪) را نشان می‌دهد. طبق نتایج به دست آمده از این جدول مدل ترکیبی Voting Classiifer جهت پیش‌بینی و طبقه‌بندی بهترین عملکرد را در بین تمامی مدل‌ها داشت. پژوهش حاضر با استفاد از دیتاست مطالعه یاس و ۳۰ پارامتر مربوطه از داده‌های یاس شامل سن، جنس، ضربان قلب فشارخون در حالت استراحت و ....) و الگوریتم‌های SVM، XGBoostو Voting Classiifer انجام شد. تجزیه و تحلیل داده‌ها با استفاده از الگوریتم Voting Classiifer دقت ۹۰٪ در پیش‌بینی بیماری ایسکمیک قلبی را نشان داد که در مقایسه باقسمت نخست نتایج این مطالعه که با استفاده از ترکیب الگوریتم‌های رندوم فارست و نایوبیز و ۲۰ پارامتر و نرم‌افزار رپیدماینر انجام شده بود دقت ۷۴/۵۱ را برای پیش‌بینی بیماری قلبی نشان داده بود (۱۵) نشان‌دهنده این است که مطالعه حاضر نتیجه مطلوب‌تر و دقت بالاتری را ارائه داد. مدل XGBClassifier برروی داده‌های آموزشی، آموزش داده شد و اهمیت ویژگی‌ها استخراج گردید.

مهم‌ترین ویژگی‌های شناسایی شده به ترتیب عبارتند از: دیابت، سابقه خانوادگی بیماری، کلسترول بالا، فشارخون بالا، سن، فشارخون دیاستولیک (DBP)، فشارخون سیستولیک (SBP) و شاخص توده‌بدنی (BMI). بر اساس این تحلیل، ۸ ویژگی برتر برای مرحله طبقه‌بندی انتخاب شدند و درنهایت سن یکی از مهمترین و تأثیرگذارترین عوامل در پیش‌بینی بیماری‌های ایسکمیک قلب در تمامی مدل‌ها بود. (نمودار ۱) مقایسه معیارها (Accuracy،Precision, Recall, F1-Score) برای SVM، XGBoostو Voting Classiifer نمایش می‌دهد این نمودار یک نمودار میله‌ای است که انتظار می‌رود که میله Voting Classiiferبرای تمام معیارها بلندتر باشد. (نمودار ۱) پنج معیار کلیدی (دقت، صحت، فراخوان، نمره F1 و ROC AUC) برای هر عملگر نمایش می‌دهد، سه میله، یکی برای هر مدل(SVM، XGBOOST، رأی دهنده کلاس). این امکان مقایسه سریع بصری و شناسایی مدل پیشرو برای هر عملگر را فراهم می‌کند. از نمودار قابل مشاهده است که طبقه‌بندی رای‌دهنده به‌طور مداوم بالاترین نوار را برای همه معیارها به‌دست می‌آورد که نشانگر برتری آن است.

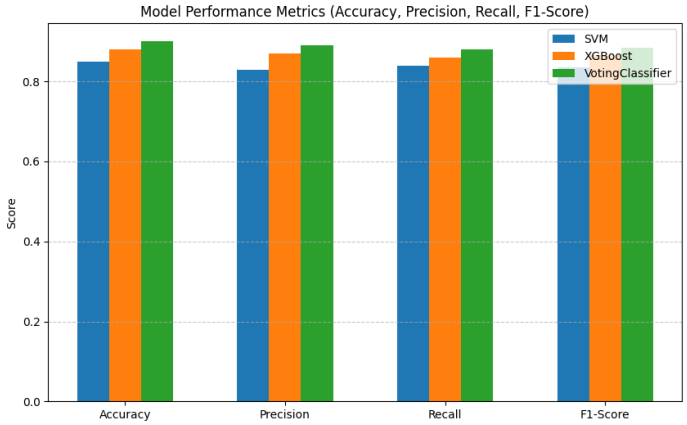
در (نمودار ۲) منحنی ROC و AUC مدل‌ها (این نمودار شامل سه منحنی ROC برای MVS، XGBoost و Voting Classiifer بر روی یک صفحه مختصات بدست آمده است. مساحت زیر هر منحنی AUC آن مدل را نشان می‌دهد. منحنی Voting Classiifer در بالا و سمت چپ قرار گرفته و دارای بیشترین مساحت زیر منحنی است که به معنی مدل در تشخیص درست بین کلاس‌ها (مثلاً بیمار و سالم) توانمندتر است.

[Downloaded from jssu.ssu.ac.ir on 2026-10-21]
[DOI: 10.18502/ssu.v34i7.22992]

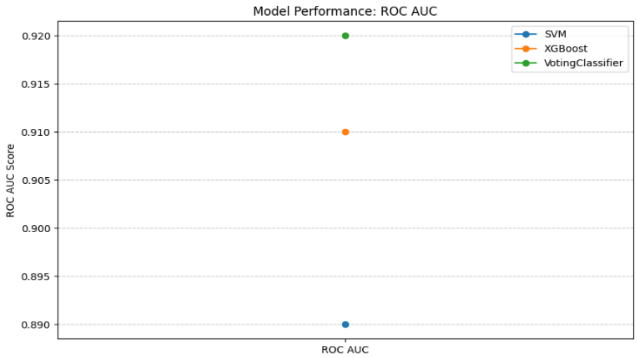
CC BY-NC 4.0

جدول۳: نتایج ارزیابی عملکرد مدل‌های SVM، XGBoost و Voting Classiifer در شرکت کنندگان فاز اول مطالعه سلامت مردم یزد ۱۳۹۴-۱۳۹۳

معیار	مدل SVM	مدل XGBoost	مدل ترکیبی Voting Classiifer
Accuracy	0.85	0.88	0.90
Precision	0.83	0.87	0.89
Recall	0.84	0.86	0.88
F1-Score	0.835	0.865	0.885
ROC AUC	0.89	0.91	0.92



نمودار۱: مقایسه عملکرد مدل‌ها در پیش‌بینی بیماری‌های ایسکمیک قلب در شرکت‌کنندگان فاز اول مطالعه سلامت مردم یزد ۱۳۹۳-۹۴



نمودار۲: منحنی ROC و AUC مدل‌ها

بحث

هدف مطالعه حاضر ارائه ابزاری کارآمد برای تشخیص زودهنگام بیماری‌های ایسکمیک قلبی و کمک به تصمیم‌گیری‌های بالینی، با استفاده از الگوریتم‌های ترکیبی می‌باشد. با پیش‌بینی و تشخیص زودهنگام می‌توان درمان‌های لازم را در مراحل اولیه برای بیماران انجام داده و باعث کاهش مرگ‌ومیر شد. Mahmoodi پژوهشی باهدف طراحی یک سیستم هوشمند برای تشخیص بیماری قلبی با استفاده از کامپیوتر انجام داد. مجموعه داده مورد استفاده ۲۷۰ بیمار مراجعه‌کننده که دارای ۱۳ ویژگی (سن، جنس، ضربان قلب، فشارخون در حالت استراحت، نوع درد قفسه‌سینه، کلسترول خون، قند خون ناشتا، نتایج الکتروکاردیوگرافی در حالت استراحت و ....) بودند توسط تکنیک فازی و الگوریتم ماشین بردار پشتیبان در نرم‌افزار متلب جهت تشخیص درست و سریع که درصد نجات بیمار را افزایش می‌دهد استفاده شد. همچنین معیارهای ارزیابی در این سیستم نرخ دسته‌بندی و حساسیت بود که عملکرد این سیستم بر اساس این شاخص‌ها به ترتیب ۸۵٪ و ۸۵/۸٪ به‌دست‌آمده بود که مدل پیشنهادی با دقت نسبتاً بالایی افراد مبتلابه بیماری ایسکمیک قلبی را تشخیص داده بود (۱۶). مطالعه Setiyadi و همکاران با هدف مقایسه عملکرد سه الگوریتم کلیدی جنگل تصادفی، ماشین

CC BY-NC 4.0
[Downloaded from jssu.ssu.ac.ir on 2026-10-21]
[DOI: 10.18502/ssu.v34i7.22992]

بردار پشتیبان و شبکه‌های عصبی در پیش‌بینی بیماری قلبی بود. با استفاده از مجموعه داده بیمار با دو ویژگی اسمی و عددی، تکنیک‌های داده کاوی از طریق نرم‌افزار Orange به‌کار گرفته شد. طبقات هدف نشان دهنده عدم وجود (۰) یا وجود (۱) اختلالات قلبی بود. نتایج نشان داد که SVM با نرخ ۸۵٪ بالاترین دقت را به‌دست آورد و عملکرد بهتری از RF و NN داشت. یافته‌ها نشان می‌دهد که الگوریتم SVM ابزاری قابل‌اعتماد برای پیش‌بینی بیماری ایسکمیک قلبی است که به کاهش خطاهای تشخیصی و بهبود تصمیم‌گیری پزشکی کمک می‌کند (۱۷). مطالعه Gandha با بررسی اثربخشی الگوریتم XGBoost، استفاده از یادگیری ماشین را برای پیش‌بینی بیماری‌های ایسکمیک قلبی بررسی می‌کند. XGBoost را در برابر Decision Trees و Random Forests با استفاده از مجموعه داده Kaggle که شامل بیش از ۱۰۰۰ پرونده بیمار با چهارده ویژگی مهم بود، آزمایش شد. با امتیاز F1 برابر ۰/۹۸۱۶ با فراخوانی ۱/۰، دقت ۹۶/۳۹٪ و دقت ۹۸/۰۴٪ مدل XGBoost بهینه شده با فراپارامتر عملکرد تحسین برانگیزی داشت. رویکردهای قبلی، مانند مدل جنگل تصادفی ترکیبی، که به دقت ۸۸/۷٪ با مجموعه داده کاهش یافته دست یافت، توسط این مدل پیشی گرفت. هیچ نمونه گم نشده و تعداد کمی از موارد مثبت کاذب وجود نداشت که نشان دهنده قابلیت اطمینان مدل XGBoost در پیش‌بینی بیماری قلبی است (۱۸). این مطالعه اثربخشی دو الگوریتم طبقه‌بندی، ماشین‌بردار پشتیبان (SVM) و K-نزدیک‌ترین همسایه (KNN) را در پیش‌بینی خطر بیماری قلبی با استفاده از مجموعه داده Kaggle متشکل از ۳۰۳ رکورد با ۱۴ ستون ویژگی مقایسه می‌کند. داده‌ها به ۷۰ برای آموزش و ۳۰ برای آزمایش تقسیم می‌شوند. نرم‌افزار مورد استفاده در این مطالعه داده‌کاوی نارنجی برای ساخت مدل‌های SVM و KNN است. نتایج نشان می‌دهد که دقت SVM ۸۵/۶٪ است، در حالی که KNN به ۸۱/۱٪ می‌رسد. بر اساس ماتریس سردرگمی، الگوریتم SVM نسبت به KNN نرخ خطای کمتری دارد. در نتیجه، الگوریتم SVM در پیش‌بینی خطر بیماری ایسکمیک قلبی نسبت به KNN برتر است (۱۹). این مطالعه با هدف ایجاد یک مدل ML با دقت بالا برای پیش‌بینی بیماری قلبی با استفاده از سه روش مجزا (مجذور کای، ANOVA و اطلاعات متقابل) برای انتخاب بهینه‌ترین ویژگی‌ها استفاده شد، که منجر به تشکیل سه زیرمجموعه (SF-1، SF-2 و SF-3) شد سپس ده طبقه‌بندی‌کننده، ماشین بردار پشتیبان ((SVM، voting، XGBoost، AdaBoost، bagging، درخت تصمیم (DT)، K نزدیک‌ترین همسایه (KNN)، جنگل تصادفی (RF) و رگرسیون لجستیک (LR) برای یافتن بهترین تکنیک و مناسب‌ترین زیرمجموعه ویژگی‌ها آزمایش شدند. برای رسیدگی به مشکل داده‌های نامتعادل، از تکنیک نمونه‌برداری بیش از حد اقلیت مصنوعی (SMOTE) استفاده شد. نتایج نشان داد که طبقه‌بندی کننده XGBoost با استفاده از مجموعه داده‌های ترکیبی و زیرمجموعه ویژگی SF-2 با نرخ‌های زیر به عملکرد بهینه دست یافته است: ۹۷/۵۷٪ برای دقت، ۹۶/۶۱٪ برای حساسیت، ۹۰/۴۸٪ برای ویژگی، ۹۵/۰۰٪ برای دقت، ۹۲/۶۸٪ برای امتیاز F1، و AUC. توسعه یک رویکرد هوش مصنوعی قابل توضیح که از متدولوژی‌های SHAP استفاده می‌کند، برای درک نحوه پیش‌بینی نتایج نهایی توسط سیستم انجام می‌شود (۲۰). Budholiya و همکاران یک سیستم تشخیصی ارائه داد که از یک طبقه‌بندی‌کننده بهینه‌شده XGBoost برای پیش‌بینی بیماری قلبی استفاده شده است. تنظیم مناسب فراپارامتر برای کاربرد موفقیت‌آمیز هر طبقه‌بندی‌کننده ضروری است. برای بهینه‌سازی فراپارامترهای XGBoost، از بهینه‌سازی بیزی استفاده شد که روشی بسیار کارآمد برای بهینه‌سازی فراپارامتر است. همچنین از تکنیک کدگذاری One-Hot (OH) برای کدگذاری ویژگی‌های دسته‌بندی‌شده در مجموعه داده‌ها برای بهبود دقت پیش‌بینی استفاده کردند. اثربخشی مدل پیشنهادی بر روی مجموعه داده‌های بیماری قلبی کلیولند Cleavland ارزیابی شده و آن را با طبقه‌بندی‌کننده‌های جنگل تصادفی و درخت اضافی مقایسه شد. نتایج تجربی مدل پیشنهادی با دقت پیش‌بینی بالای ۹۱/۸٪ عملکرد بهتری را در مقایسه با مدل‌های پیشنهادی قبلی داشته است (۲۱). مطالعه حاضر نشان می‌دهد ترکیب الگوریتم‌هایی مانند SVM و XGBoost یا استفاده از الگوریتم‌های Ensemble دیگر، منجر به نتایج بهبود یافته در حوزه سلامت شده است. مقایسه معیارهای ارزیابی (مانند Accuracy و F1-Score) با مدل‌هایی که در سایر پژوهش‌ها بر روی مجموعه داده‌های مشابه یا متفاوت مورد استفاده قرار گرفته‌اند، نشان‌دهنده جایگاه رقابتی این مدل در میان روش‌های موجود است.

محدودیت‌ها و چالش‌ها: با وجود نتایج امیدوارکننده، این تحقیق با محدودیت‌هایی نیز روبرو است:

محدودیت مجموعه داده: عملکرد مدل به شدت وابسته به کیفیت و کمیت داده‌های موجود است. اگر مجموعه داده های یاس دارای

CC BY-NC 4.0

[Downloaded from jssu.ssu.ac.ir on 2026-10-21]
[DOI: 10.18502/ssu.v34i7.22992]

واریانس کم یا سوگیری خاصی باشد، این موارد ممکن است بر قابلیت تعمیم‌پذیری مدل به جمعیت‌های دیگر تأثیر بگذارند.

قابل تفسیر بودن(Interpretability): در حالی‌که XGBoost تا حدی قابلیت تفسیرپذیری را از طریق اهمیت ویژگی‌ها ارائه می‌دهد، مدل‌های Ensemble مانند Voting Classiifer و هم‌چنین (SVMبه‌خصوص با کرنل‌های پیچیده) ممکن است کمتر قابل تفسیر باشند. درک دلیل دقیق اینکه چرا مدل یک فرد را در معرض خطر تشخیص می‌دهد، برای پزشکان بسیارمهم است.

انتقال‌پذیری (Transferability): مدل توسعه‌یافته بر اساس داده‌های خاص شهرستان یزد است. ممکن است نیاز به تنظیم یا آموزش مجدد برای استفاده در مناطق یا جمعیت‌های دیگر با ویژگی‌های دموگرافیک و ژنتیکی متفاوت وجود داشته باشد.

نتیجه‌گیری

در این پژوهش، یک مدل ترکیبی یادگیری ماشین با استفاده از الگوریتم‌های SVM و XGBoost با رویکرد VotingClassifier، برای پیش‌بینی بیماری‌های ایسکمیک قلبی با استفاده از مجموعه داده‌های فاز اول (ثبت نام) مطالعه سلامت مردم یزد (یاس) انجام شد. مراحل پیش پردازش دقیق داده‌ها، شامل مدیریت مقادیر گمشده، استانداردسازی و تبدیل ویژگی‌های دست‌های، به منظور آماده‌سازی داده‌ها برای آموزش مدل‌ها انجام شد. نتایج حاصل از ارزیابی مدل‌ها بر روی مجموعه آزمون نشان داد که مدل ترکیبی VotingClassiifer عملکرد بهتری نسبت به مدل‌های منفرد SVM و XGBoost در معیارهای Accuracy, Precision, Recall, F1-Scoreو ROC AUC از خود نشان می‌دهد. این برتری، اثربخشی رویکرد Ensemble در ترکیب نقاط قوت مدل‌های پایه و کاهش خطاهای پیش‌بینی را تأیید می‌کند. دستیابی به مقادیر بالای معیارهای ارزیابی (مانند ROC AUC 0.92) پتانسیل قابل‌توجه این مدل را برای استفاده در تشخیص زودهنگام بیماری ایسکمیک قلبی آشکار می‌سازد. تشخیص زودهنگام و دقیق بیماری ایسکمیک قلبی از اهمیت بالایی برخوردار است و مدل توسعه یافته می‌تواند به‌عنوان ابزاری کارآمد در اختیار پزشکان و متخصصان بهداشت قرار گیرد تا با شناسایی افراد در معرض خطر، امکان مداخله به موقع، کاهش مرگ‌ومیر و بهبود کیفیت زندگی بیماران را فراهم آورد. با وجود محدودیت‌هایی مانند قابلیت تفسیرپذیری و نیاز به اعتبارسنجی بر روی داده‌های متنوع‌تر، این تحقیق گام مهمی در جهت استفاده از فناوری‌های نوین یادگیری ماشین برای ارتقاء سلامت جامعه، به‌ویژه در حوزه پیش‌بینی و پیشگیری از بیماری‌های ایسکمیک قلبی، محسوب می‌شود. تحقیقات آینده می‌تواند با بررسی الگوریتم‌های Ensemble پیشرفته‌تر، روش‌های عمیق‌تر تفسیرپذیری و استفاده از مجموعه داده‌های وسیع‌تر، این زمینه را بیش از پیش توسعه دهد.

پیشنهادات آینده‌نگر

در زیر، پیشنهادات آینده‌نگر و استراتژیک برای فاز بعدی پروژه مدل‌سازی داده‌های یاس(YaHS) آورده شده است: پیشنهاد می‌شود در پژوهش‌های آتی از معماری‌های یادگیری عمیق (Deep Learning) و شبکه‌های عصبی (CNN/RNN/Transformer) استفاده گردد.

سپاس‌گزاری

این مقاله بخشی از پایان‌نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش نرم‌افزار دانشگاه پیام نور تهران می‌باشد که بدون حمایت مالی انجام شده است. در پایان از تمامی شرکت‌کنندگان و مجریان طرح یاس که امکان انجام این تحقیق را فراهم نموده‌اند، تشکر می‌گردد.

حامی مالی

ندارد.

تعارض در منافع

وجود ندارد.

ملاحظات اخلاقی

پروپوزال این تحقیق توسط کمیته اخلاق دانشگاه علوم پزشکی شهید صدوقی یزد تایید شده است (کداخلاق: IR.SSU.REC.1401.016).

اظهارنامه استفاده از فناوری هوش مصنوعی

در نگارش این مقاله از هیچ ابزار یا فناوری هوش مصنوعی استفاده نشده است.

دسترس‌پذیری داده ها

داده های خام این مطالعه از طریق درخواست از نویسنده مسئول ارائه خواهد شد.

مشارکت نویسندگان

مسعود میرزایی مجری اول و پایه گذار مطالعه یاس است. اعظم برزگری، مهیار عبدالملکی، مسعود میرزایی در ارائه ایده، اعظم برزگری، مهیار عبدالملکی در طراحی مطالعه، اعظم برزگری در جمع‌آوری داده‌ها، مهیار عبدالملکی در تجزیه و تحلیل داده‌ها مشارکت داشته و همه نویسندگان در تدوین، ویرایش اولیه و نهایی مقاله و پاسخگویی به سوالات مرتبط با مقاله سهیم هستند.

CC BY-NC 4.0

[Downloaded from jssu.ssu.ac.ir on 2026-10-21]
[DOI: 10.18502/ssu.v34i7.22992]

References

1. Safdari R, Ghazi Saeedi M, Gharooni M, Nasiri M, Argi G. ***Comparing performance of decision tree and neural network in predicting myocardial infarction.*** J Paramed Sci Rehabil. 2014; 3(2): 26-35.

2. Premsmith J, Ketmaneechairat H. ***A predictive model for heart disease detection using data mining techniques.*** J Adv Inf Technol. 2021; 12(1): 14-20.

3. Fasihi L, Siahkohian M, Ebrahimi-Torkamani B. ***Using support vector machine algorithm to predict coronary heart disease in active middle-aged women.*** J Mil Med. 2023; 25(5): 2016-2023.

4. Siahkohian M, Fasihi L, Ebrahimi Torkamani B. ***Prediction of coronary heart disease using discriminant analysis algorithm in active elderly men.*** J Ardabil Univ Med Sci. 2023; 22(4): 370-379.

5. Tabatabaei SM, Saadatjoo F, Mirzaei M. *The prediction model for cardiovascular disease using Yazd's health study data (YaHS)*. J Shahid Sadoughi Univ Med Sci. 2019; 27(3): 1346-1360.

6. Sianga BE, Mbago MC, Msengwa AS. ***Predicting the prevalence of cardiovascular diseases using machine learning algorithms.*** Intell Based Med. 2025; 11: 100199.

7. Ozcan M, Peker S. ***A classification and regression tree algorithm for heart disease modeling and prediction.*** Healthc Anal. 2023; 3: 100130.

8. Mirzaei M, Salehi-Abargouei A, Mirzaei M, Mohsenpour MA. ***Cohort profile: the Yazd Health Study (YaHS): a population-based study of adults aged 20-70 years (study design and baseline population data).*** Int J Epidemiol. 2018; 47(3): 697-698h.

9. Chandra MA, Bedi SS. ***Survey on SVM and their application in image classification.*** Int J Inf Technol. 2021; 13(5): 1-11.

10. Ramraj S, Uzir N, Sunil R, Banerjee S. ***Experimenting XGBoost algorithm for prediction and classification of different datasets.*** Int J Control Theory Appl. 2016; 9(40): 651-662.

11. Cabena P, Hadjinian P, Stadler R, Verhees J, Zanasi A. ***Discovering data mining: from concept to implementation.*** Prentice-Hall; 1998.

12. Jalaly SH, Abdolvand N, Rajaee Harandi S. ***Customer behavior analysis using web usage mining.*** Iran Mark Sci Assoc. 1397; 8(2): 69-86.

13. Bishop CM, Nasrabadi NM. ***Pattern recognition and machine learning.*** New York: Springer; 2006.

14. Hyppönen V. ***A comparative analysis of migration tools for Python 3 in medical device software.*** 2025.

15. Barzegari A, Noorani SF, Mirzaei M. ***Comparison of data mining algorithms in prediction of coronary artery diseases using Yazd Health Study (YaHS) data.*** J Shahid Sadoughi Univ Med Sci. 2023; 31(7): 6824-6835.

16. Mahmoodi MS. ***Designing a heart disease prediction system using support vector machine.*** J Health Biomed Inform. 2017; 4(1): 1-10.

17. Setiyadi D, Henderi H, Suryaningrat A, Swastika R, Saludin S, Mutoffar MM, et al. ***Prediction of heart disease using random forest algorithm, support vector machine, and neural network.*** Telkomnika. 2025; 23(1): 129-137.

18. Gandha R, Richhariya P. ***XGBoost for heart disease prediction: achieving high accuracy with robust machine learning techniques.*** Int J Innov Sci Eng Manag. 2024; 11(7): 07-14.

19. Pangaribuan JJ, Maulana A, Romindo R. ***Unleashing the power of SVM and KNN: enhanced early detection of heart disease.*** JITK. 2024; 10(2): 342-351.

20. El-Sofany HF. ***Predicting heart diseases using machine learning and different data classification techniques.*** IEEE Access. 2024; 12: 106846-106859.

21. Budholiya K, Shrivastava SK, Sharma V. ***An optimized XGBoost based diagnostic system for effective prediction of heart disease.*** J King Saud Univ Comput Inf Sci. 2022; 34(7): 4514-4523.

[Downloaded from jssu.ssu.ac.ir on 2026-10-21]

[DOI: 10.18502/ssu.v34i7.22992]

CC BY-NC 4.0