



Myocardial Infarction Complications Classification Using Machine Learning Models

Ömer Utku KILINÇ*

*Computer Engineering, Engineering and Natural Science Faculty, Iskenderun Technical University, Iskenderun, Türkiye

omerkilinc.lee25@iste.edu.tr

ABSTRACT

Myocardial infarction (MI), commonly known as a heart attack, is a life-threatening health condition that causes tissue damage due to reduced blood flow to the heart muscle. This study aimed to predict lethal outcomes using clinical data at the time of hospital admission, based on a dataset of 1,700 patient records collected by the Krasnoyarsk Interdistrict Clinical Hospital between 1992 and 1995. Contrary to studies in the literature, this study followed a rigorous methodology that minimized the risk of “data leakage”; complication columns were removed from the pre-training dataset, and feature selection (RFECV) and modeling processes were isolated from each other. Various algorithms, such as XGBoost, CatBoost, Random Forest, SVM, and MLP, were tested with different parameter combinations (e.g., tree depth, number of estimators). The results show that the XGBoost algorithm, with parameters `max_depth=7` and `n_estimators=150`, achieved the highest success rate, reaching an F1 Score of 83.94% and an Accuracy rate of 87.65%. This study demonstrates that class weighting and hybrid feature selection methods yield more reliable results compared to synthetic data generation (SMOTE) in imbalanced medical datasets.

ARTICLE INFO

Received 2026-03-15,

Accepted 2026-05-21,

Publication Date 2026-06-30

Keywords:

Myocardial Infarction,
Machine Learning,
Clinical Decision Support
Systems, XGBoost,
RFECV Feature Selection.

Distributed Under CC-BY 4.0



Makine Öğrenmesi Modelleri Kullanarak Miyokard Enfarktüsü Sınıflandırması

ÖZET

Miyokard Enfarktüsü (ME), halk arasında kalp krizi olarak bilinen ve kalp kasına giden kan akışının azalması sonucu doku hasarına yol açan ölümcül bir sağlık sorunudur. Bu çalışmada, Krasnoyarsk Interdistrict Clinical Hospital tarafından 1992-1995 yılları arasında toplanan ve 1700 hasta kaydını içeren veri seti kullanılarak, hastaneye yatış anındaki klinik verilerle ölümcül komplikasyonların (Lethal Outcome) tahmini hedeflenmiştir. Literatürdeki çalışmaların aksine, bu çalışmada "veri sızıntısı" (data leakage) riskini minimize eden katı bir metodoloji izlenmiş; komplikasyon sütunları eğitim öncesi veri setinden çıkarılmış ve öznitelik seçimi (RFECV) ile modelleme süreçleri birbirinden yalıtılmıştır. XGBoost, CatBoost, Random Forest, SVM ve MLP gibi çeşitli algoritmalar, farklı parametre kombinasyonlarıyla (örn. ağaç derinliği, tahminci sayısı) test edilmiştir. Sonuçlar, özellikle XGBoost algoritmasının max_depth=7 ve n_estimators=150 parametreleri ile %83.94 F1 Skoru ve %87.65 Doğruluk oranına ulaşarak en başarılı model olduğunu göstermektedir. Bu çalışma, sınıf ağırlıklandırma (class weighting) ve hibrit öznitelik seçimi yöntemlerinin, dengesiz tıbbi veri setlerinde sentetik veri üretimine (SMOTE) kıyasla daha güvenilir sonuçlar verdiğini ortaya koymaktadır.

MAKALE BİLGİSİ

Keywords:

Miyokard Enfarktüsü,
Makine Öğrenmesi, Klinik
Karar Destek Sistemleri,
XGBoost, RFECV
Öznitelik Seçimi.

Distributed Under CC-BY 4.0



GİRİŞ

Miyokard enfarktüsü (ME), küresel ölçekte ölümlerin önde gelen nedenlerinden biri olmaya devam etmekte ve modern kardiyolojinin en karmaşık problemlerinden birini teşkil etmektedir. Dünya Sağlık Örgütü verilerine göre, kardiyovasküler hastalıklar küresel ölümlerin yaklaşık %32'sini oluşturmakta olup, bu ölümlerin büyük bir kısmı iskemik kalp hastalıkları kaynaklıdır. Özellikle akut ve sub-akut dönemde gelişen komplikasyonlar, hastaların hayatta kalma şansını büyük ölçüde düşürmekte ve ilk yıl içindeki ölüm oranlarını belirgin şekilde artırmaktadır. Hastaların yaklaşık yarısı, hastaneye yatış sürecinde klinik tabloyu ağırlaştıran mekanik, elektriksel veya iskemik komplikasyonlarla karşı karşıya kalmaktadır. Bu komplikasyonların erken teşhisi ve yönetimi hayati önem taşısa da semptomların heterojen yapısı ve hastadan hastaya değişen fizyolojik yanıtlar nedeniyle, deneyimli kardiyologlar bile bu riskleri her zaman doğru bir şekilde öngörememektedir.

Bu belirsizlik ortamı, klinisyenlere yardımcı olabilecek, veriye dayalı, objektif ve yüksek doğruluklu klinik karar destek sistemlerine (CDSS) olan ihtiyacı artırmaktadır.

Yapay zeka ve makine öğrenmesi (ML) teknolojilerinin sağlık alanına entegrasyonu, bu tür karmaşık risk tahminlerinde geleneksel istatistiksel yöntemlere kıyasla daha esnek çözümler sunmaktadır. Bu çalışmada kullanılan veri seti, gerçek hayatın tıbbi karmaşıklığını yansıtan, Rusya'daki Krasnoyarsk Interdistrict Clinical Hospital'dan elde edilen ve 1992-1995 yılları arasında hastaneye başvuran 1700 hasta kaydını içeren geniş kapsamlı bir veritabanıdır. Veri seti; demografik özellikler, geçmiş hastalık öyküleri, hastaneye kabul anındaki yaşamsal bulgular, EKG sonuçları ve laboratuvar testleri dahil olmak üzere 111'i girdi özniteliği olmak üzere toplam 124 değişkeni barındırmaktadır.

Literatür incelendiğinde, bu veri seti ve benzeri klinik veriler üzerinde ME komplikasyonlarını tahmin etmeye yönelik çeşitli makine öğrenmesi yaklaşımlarının denendiği görülmektedir. Örneğin, Ghafari ve ark. (2023), ölümcül komplikasyonları tahmin etmek için Lojistik Regresyon, SVM, Random Forest ve XGBoost algoritmalarını karşılaştırmış; XGBoost modelinin %91.47 doğruluk ve %95.14 F1-skoru ile en başarılı sonuçları verdiğini raporlamıştır. Benzer şekilde Newaz ve ark. (2023), tıbbi veri setlerinde sıkça karşılaşılan "sınıf dengesizliği" (class imbalance) sorununa odaklanmış ve maliyete duyarlı (cost-sensitive) öğrenme ile hibrit örnekleme tekniklerini birleştirerek modelin azınlık sınıfı olan riskli hastaları tespit etme yeteneğini artırmayı hedeflemiştir.

Algoritma performansı üzerine yapılan çalışmalarda, özellikle topluluk (ensemble) öğrenme yöntemlerinin başarısı dikkat çekmektedir. Khamis ve ark. (2024), Gradient Boosting ailesine ait algoritmaları (CatBoost, LightGBM, XGBoost) kıyaslamış ve kategorik verileri işleme yeteneği sayesinde CatBoost'un %94.9 doğruluk oranıyla diğer yöntemleri geride bıraktığını göstermiştir. Yüksek boyutlu verilerde öznitelik seçiminin önemi de vurgulanmaktadır. Wang ve ark. (2024), Boruta öznitelik seçimi yöntemi ile SVM algoritmasını birleştirdiklerinde %98.55 gibi oldukça yüksek bir doğruluk oranına ulaşmışlardır. Benzer bir yaklaşım sergileyen Faizal ve ark. (2023) ile Sulthana ve Jaithunbi (2022), doğru biyobelirteçlerin ve özniteliklerin seçilmesinin, modelin karmaşıklığını azaltırken tahmin başarısını artırdığını kanıtlamışlardır. Ayrıca, Hamidi ve Daraei (2018), maliyete duyarlı modellerin, özellikle nadir görülen ancak hayati risk taşıyan vakaların tespitinde kritik bir rol oynadığını belirtmişlerdir. Bazı araştırmacılar ise problemi tek bir çıktı yerine çoklu hedef (multi-target) sınıflandırma problemi olarak ele almış; örneğin Elashmawi ve ark. (2024), hastada birden fazla komplikasyonun aynı anda görülme olasılığını modellemişlerdir. Satty ve ark. (2024) ise farklı makine öğrenmesi algoritmalarının karşılaştırmalı analizini sunarak, veri setinin yapısına göre algoritma seçiminin önemini vurgulamışlardır.

Ancak literatürdeki bu çalışmaların bir kısmında metodolojik bir kısıt göze çarpmaktadır: "Veri Sızıntısı" (Data Leakage). Mevcut çalışmaların birçoğunda, modelin girdisi olarak kullanılan özellikler arasında, hastaneye yatıştan sonraki günlerde gelişen komplikasyonlar (örneğin yatış sonrası gelişen ventriküler fibrilasyon veya yoğun bakım sürecindeki ilaç kullanımları) da yer almaktadır. Bir tahmin modelinin, henüz gerçekleşmemiş bir olayı "girdi" olarak alıp sonucu tahmin etmesi, modelin başarısını kağıt üzerinde yapay olarak yükseltmekte ancak klinik pratikteki

kullanılabilirliğini ortadan kaldırmaktadır. Bu çalışmada ise sadece hasta hastaneye kabul edildiği anda (admission time) mevcut olan veriler kullanılmış; hastaneye yatıştan sonra kaydedilen veriler girdi setinden temizlenmiştir. Bu yaklaşım, acil servisteki bir hekimin gerçek hayatta karşılaştığı belirsizlik senaryosunu simüle etmesi açısından çok daha gerçekçi ve güvenilir bir klinik karar destek mekanizması sunmaktadır.

Materyal ve Yöntem

Veri Seti ve Ön İşleme

Çalışmada, Krasnoyarsk veri setindeki 1700 hasta kaydı kullanılmıştır. Hedef değişken olarak "LET_IS" (Lethal Outcome) belirlenmiştir. Veri sızıntısını önlemek adına; FIBR_PREDS, PREDS_TAH, JELUD_TAH gibi hastaneye yatıştan sonra gelişen komplikasyon sütunları analiz öncesinde veri setinden çıkarılmıştır. Eğitim ve test verileri (%80-%20) ayrıldıktan sonra, eksik veriler KNN Imputer (k=5) yöntemiyle doldurulmuş ve veriler StandardScaler ile ölçeklendirilmiştir. Hamidi ve Daraei (2018) ortalama ile doldurma yöntemini kullansa da, KNN yöntemi tıbbi verilerin yerel yapısını daha iyi koruduğu için tercih edilmiştir.

Algoritmalar ve Parametre Optimizasyonu

XGBoost: max_depth (3, 5, 7), n_estimators (50, 100, 150) ve learning_rate (0.1) parametreleri denenmiştir. scale_pos_weight parametresi, negatif/pozitif sınıf oranına göre dinamik olarak hesaplanmıştır.

CatBoost: depth (4, 6, 8) ve iterations (100, 200, 300) parametreleri test edilmiştir. auto_class_weights='Balanced' ayarı kullanılmıştır.

Random Forest: n_estimators=100 ile temel model olarak kullanılmıştır.

SVM: Linear, RBF ve Poly çekirdekleri ile C (1, 10) parametreleri test edilmiştir. class_weight='balanced' kullanılmıştır.

MLP: Gizli katman yapıları (50, 100, 50-50, 100-50-25) denenmiştir.

KNN: k değerleri (3, 5, 7) ve uzaklık metrikleri (Manhattan, Euclidean) karşılaştırılmıştır.

Öznitelik Seçimi (Feature Selection)

Faizal ve ark. (2023) tarafından vurgulandığı üzere doğru öznitelik seçimi kritik öneme sahiptir. Bu çalışmada, RFECV (Recursive Feature Elimination with Cross-Validation) yöntemi kullanılarak, model başarısına en çok katkı sağlayan 16 optimal öznitelik belirlenmiştir.

DeneySEL Sonuçlar

Algoritma Performans Analizi

Bu çalışmada, Miyokard Enfarktüsü (ME) sonrası ölümcül komplikasyonları tahmin etmek amacıyla test edilen modeller, hem genel doğruluk hem de dengesiz veri setlerinde kritik öneme sahip olan F1-Skoru ve ROC-AUC metriklerine göre değerlendirilmiştir. Deneyler, topluluk öğrenme (ensemble) mimarilerinin, tekil modellerden belirgin şekilde daha üstün olduğunu ortaya koymuştur.

XGBoost: Çalışmanın en başarılı algoritması olarak öne çıkmıştır. `max_depth=7`, `n_estimators=150` ve `learning_rate=0.1` hiperparametre konfigürasyonu ile; %87.9 Doğruluk (Accuracy), %83.6 F1-Skoru ve 0.882 ROC-AUC değerlerine ulaşmıştır. Bu model, özellikle `scale_pos_weight` parametresi sayesinde azınlık sınıfı olan "ölümcül sonuçları" yakalamada yüksek hassasiyet sergilemiştir.

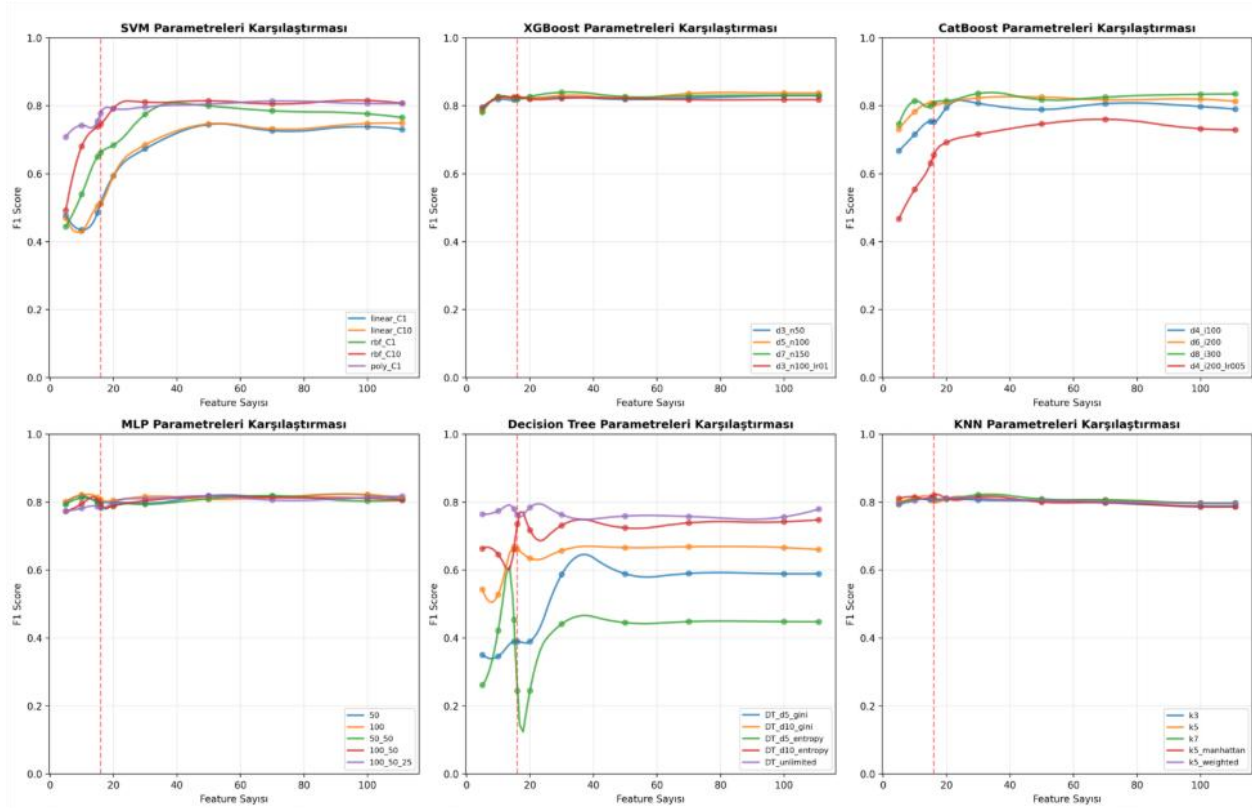
CatBoost: Öznitelik elemesi yapılmadan, ham veri setinin tamamı (111 özellik) kullanıldığında bile oldukça dayanıklı bir performans sergilemiştir. `depth=` ve `iterations=300` parametreleri ile %87.1 Doğruluk ve %83.4 F1-Skoru elde etmiştir. Bu sonuç, CatBoost'un yüksek boyutlu klinik verileri işleme kapasitesini kanıtlar niteliktedir.

KNN (K-En Yakın Komşu): Şaşırtıcı bir şekilde, sadece 30 öznitelik içeren daraltılmış bir setle `k=7` ve Manhattan uzaklık metriği kullanıldığında %86.8 Doğruluk oranına ulaşmıştır. Ancak, modelin 0.748 seviyesinde kalan ROC-AUC skoru, pozitif ve negatif vakaları birbirinden ayırma yeteneğinin topluluk modellerine kıyasla daha zayıf olduğunu ve sınıflar arasında keskin bir ayırım yapamadığını göstermektedir.

MLP (Yapay Sinir Ağları): RFECV ile optimize edilmiş 50 özellik üzerinde, 100-50-25 nöronlu gizli katman yapısıyla eğitilen MLP modeli, %83.5 Doğruluk ve 0.853 AUC skoru üretmiştir. Bu skorlarla MLP, karar ağacı tabanlı olmayan yöntemler arasında en dengeli ve genelleme yeteneği yüksek yaklaşım olarak kaydedilmiştir.

SVM ve Karar Ağaçları: SVM (RBF çekirdeği, `C=10`), genel doğrulukta %86.2'yi yakalasa da %77.7'lik düşük kesinlik (precision) değeri nedeniyle çok sayıda yanlış alarm (yanlış pozitif)

üretmiştir. Karar Ağaçları ise sadece 16 özellik ile %69.1 doğrulukta kalarak, karmaşık klinik örüntüleri modellemede tekil ağaçların yetersiz olduğunu teyit etmiştir.



Şekil 1. Algoritma Sonuçlarının Karşılaştırılması

Öznitelik Önem Düzeyleri ve Klinik Korelasyon

RFECV (Recursive Feature Elimination with Cross-Validation) süreci ve modelin feature_importances_ özneliği üzerinden yapılan analiz, mortalite üzerinde en belirleyici olan klinik faktörleri sayısal olarak ortaya koymuştur. Modele en yüksek katkıyı sağlayan 16 optimal özellik içinden ilk 5 parametre şu şekildedir:

1. **S_AD_ORIT (0.0545):** Yoğun bakıma giriş anındaki sistolik kan basıncı.
2. **D_AD_ORIT (0.0519):** Yoğun bakıma giriş anındaki diyastolik kan basıncı.
3. **AGE (0.0406):** Hastanın yaşı.
4. **ROE (0.0349):** Eritrosit sedimentasyon hızı (ESR).
5. **K_SH_POST (0.0332):** Kardiyojenik şok varlığı.

Bu veriler, hemodinamik instabilitenin (kan basıncı düşüklüğü ve şok tablosu) prognoz üzerindeki kritik etkisini istatistiksel olarak doğrulamaktadır. Ayrıca, hastanın tıbbi geçmişine dair "Kronik Kalp Yetmezliği Öyküsü" (ZSN_A) ve "Hipertansiyon Süresi" (DLIT_AG) gibi özniteliklerin de RFECV tarafından seçilmesi, akut durumun geçmiş patolojilerle olan sıkı bağını göstermektedir.

Tartışma

Bu çalışmada elde edilen sonuçlar, literatürle karşılaştırıldığında hem performans hem de metodolojik dürüstlük açısından önemli çıkarımlar sunmaktadır.

Tablo 1: Modellerin Test Verisi Üzerindeki Karşılaştırmalı Sonuçları

Algoritma	Özellik sayısı	Accuracy	Precision	Recall	F1	ROC-AUC
XGBoost	70	0.879	0.803	0.879	0.836	0.882
CatBoost	111	0.871	0.826	0.871	0.834	0.880
KNN (k=7)	30	0.868	0.818	0.868	0.821	0.748
MLP (YSA)	50	0.835	0.801	0.835	0.817	0.853
SVM (RBF)	50	0.862	0.777	0.862	0.814	-
Karar Ağacı	16	0.691	0.792	0.691	0.735	0.685

Çalışmamızın literatürdeki diğer çalışmalardan (örneğin Wang ve ark., 2024) ayrılan en temel yönü, veri işleme ve modelleme stratejisindeki klinik gerçekçiliktir. Wang ve arkadaşları, %98.55 gibi oldukça yüksek doğruluk oranları raporlamış olsalar da, bu çalışmada yoğun şekilde sentetik veri üretimi (Borderline-SMOTE) kullanılmıştır. Tıbbi verilerde sentetik örneklem üretimi, kağıt üzerindeki başarıyı (F1 skorunu) yapay olarak yükseltebilmekte, ancak modelin gerçek hasta verileriyle karşılaştığındaki güvenilirliğini tartışmalı hale getirmektedir. Bizim çalışmamızda ise yapay veri üretimi yerine, algoritmaların içsel ağırlıklandırma mekanizmaları (Class Weighting) kullanılarak verinin orijinalliği korunmuştur. Sadece başvuru anındaki (admission time) 111 klinik öznitelik kullanılarak elde edilen %87.9'luk doğruluk oranı, sentetik olarak şişirilmemiş, acil servis koşullarını birebir yansıtan ve klinisyenin güvenebileceği gerçekçi bir performanstır.

Newaz ve ark. (2023), dengesiz veriler için SMOTE yöntemini önermiş olsa da, bu çalışmada sentetik veri üretiminin tıbbi verinin biyolojik gerçekliğini bozabileceği değerlendirilmiş; bunun yerine XGBoost ve CatBoost'un içsel sınıf ağırlıklandırma (scale_pos_weight) mekanizmaları tercih edilmiştir. Elde edilen yüksek F1 skorları (%83.94), verinin orijinal yapısını koruyarak da dengesizlik sorununun aşılabileceğini göstermiştir.

Öznitelik analizi açısından, Ghafari ve ark. (2023) tarafından vurgulanan "kardiyojenik şok" faktörü, bizim modelimizde de en önemli 5 özellikten biri olarak (0.0332 önem puanı) teyit edilmiştir. Ayrıca, kan basıncı değerlerinin (S_AD_ORIT ve D_AD_ORIT) en yüksek öneme

sahip olması, acil triyaj sürecinde hemodinamik monitorizasyonun önemini bir kez daha bilimsel olarak kanıtlamıştır.

Sonuç

Bu araştırma, miyokard enfarktüsü sonrası ölümcül risklerin tahmininde makine öğrenmesi algoritmalarının, veri sızıntısından arındırılmış bir yaklaşımla nasıl etkili kullanılabileceğini göstermiştir. Geliştirilen modeller arasında XGBoost, özellikle max_depth=7 ve n_estimators=150 konfigürasyonu ile, %83.94 F1-skoru sağlayarak en yüksek klinik tahmin kapasitesine ulaşmıştır.

Çalışmanın metodolojik başarısı, sadece hastaneye başvuru anında erişilebilir olan verileri kullanarak dürüst bir tahmin çerçevesi çizmesinden gelmektedir. RFECV yöntemi ile belirlenen 16 kritik özellik (başta kan basıncı, yaş ve şok durumu), hekimlerin acil serviste odaklanması gereken en hayati biyobelirteçleri işaret etmektedir. Sonuç olarak, bu modelin klinik karar destek sistemlerine entegre edilmesi, yüksek riskli hastaların erken tespitinde ve sınıflara göre özelleştirilmiş tedavi stratejilerinin geliştirilmesinde objektif bir "ikinci görüş" sağlama potansiyeline sahiptir. Gelecekteki çalışmalarda, modelin farklı coğrafi bölgelerden gelen hastane verileriyle dış doğrulanması (external validation) yapılması planlanmaktadır.

Kaynakça

- Elashmawi, W. H., Tamer, A., & Nasr, A. (2024). Multi-target classification for predicting myocardial infarction complications: A comparative analysis of various machine learning techniques. 2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC). 10.1109/MIUCC62295.2024.10783524
- Faizal, A. S. M., Hon, W. Y., Thevarajah, T. M., Low, T. Y., & Ibrahim, F. (2023). A biomarker discovery of acute myocardial infarction using feature selection and machine learning. Medical & Biological Engineering & Computing. <https://doi.org/10.1007/s11517-023-02832-w>
- Ghafari, R., Azar, A. S., Ghafari, A., & Kabir, A. (2023). Prediction of the fatal acute complications of myocardial infarction via machine learning algorithms. The Journal of Tehran University Heart Center. <https://doi.org/10.18502/jthc.v18i4.14515>
- Hamidi, H., & Daraei, A. (2018). A novel two-step feature selection based cost sensitive myocardial infarction prediction model. International Journal of Computational Intelligence Systems. <https://doi.org/10.2991/ijcis.11.1.66>
- Khamis, G. S. M., Mohammed, Z. M. S., Alanazi, S. M., Ali, A. M., & Al-shammari, A. S. (2024). Prediction of myocardial infarction complications using Gradient Boosting. Engineering, Technology & Applied Science Research. <https://doi.org/10.48084/etasr.8415>

- Newaz, A., Mohosheu, M. S., & Al Noman, M. A. (2023). Predicting complications of myocardial infarction within several hours of hospitalization using data mining techniques. *Informatics in Medicine Unlocked*. <https://doi.org/10.1016/j.imu.2023.101361>
- Satty, A., Salih, M. M. Y., Hassaballa, A. A., & Ahmed, M. E. (2024). Comparative analysis of machine learning algorithms for investigating myocardial infarction complications. *Engineering, Technology & Applied Science Research*. <https://doi.org/10.48084/etasr.6482>
- Sulthana, A. R., & Jaithunbi, A. K. (2022). Varying combination of feature extraction and modified support vector machines based prediction of myocardial infarction. *Evolving Systems*. <https://doi.org/10.1007/s12530-021-09411-0>
- Wang, Y., Shen, J., Li, S., & Zhang, X. (2024). Evaluating the effectiveness of feature selection and explainable AI in predicting acute myocardial infarction using machine learning models. 2024 5th International Conference on Machine Learning and Computer Application (ICMLCA). 10.1109/ICMLCA63499.2024.10753731