تقرير مشروع التنبؤ بأمراض القلب باستخدام تقنيات الذكاء الاصطناعي

1. ملخص تنفيذي

OpenML Heart، مع إضافة نماذج هجينة، وتحليل اختيار الخصائص، ونموذج Feature Tokenizer Transformer، وتفسير النتائج باستخدام SHAP.
أولية على بيانات Kaggle Cardiovascular Disease، ثم تم توسيعه ليشمل بيانات UCI Heart Disease وبيانات Disease Comprehensive 1190
يهدف هذا العمل إلى بناء إطار تجريبي قوي للتنبؤ بأمراض القلب اعتمادا على بيانات طبية منظمة ونماذج تعلم آلي متعددة. بدأ العمل من تجربة

الدقة في التحقق المتقاطع المتكرر كانت لنموذج Tuned CatBoostClassifier بدقة 94.26%، وقيمة F1 تساوي 94.56%، وROC-AUC تساوي 96.78%.
أفضل نتيجة من حيث

أفضل نتيجة من حيث ROC-AUC كانت لنموذج Tuned Soft Voting Hybrid بقيمة 97.21%، وPR-AUC تساوي 96.63%.

في اختبار holdout النهائي، حقق نموذج Tuned Stacking Hybrid قيمة ROC-AUC قدرها 97.40%، مع دقة 92.86% وقيمة F1 تساوي 93.12%.

2. هدف الدراسة

الهدف الأساسي هو تطوير نظام قابل للتكرار للتنبؤ بخطر الإصابة بأمراض القلب، مع التركيز على:

- مقارنة أكثر من مصدر بيانات بدل الاعتماد على مجموعة بيانات واحدة.
- اختبار نماذج تقليدية ونماذج boosting ونماذج هجينة.
- تفسير قرارات النموذج باستخدام SHAP.
- دراسة أثر اختيار الخصائص على الأداء.
- إضافة نموذج عميق حديث للبيانات الجدولية باستخدام Feature Tokenizer Transformer.
- تقديم نتائج قابلة للدفاع العلمي من خلال repeated stratified cross-validation واختبار holdout.

3. البيانات المستخدمة

3.1 بيانات Kaggle Cardiovascular Disease

الطول، الوزن، ضغط الدم، الكوليسترول، الجلوكوز، التدخين، النشاط البدني، والهدف cardio. تم استخدامها كبداية للتجربة ولتقييم حدود الأداء على البيانات الأصلية.
تحتوي هذه البيانات على حوالي 70 ألف سجل قبل التنظيف، وتشمل خصائص مثل العمر،

هذه البيانات، رغم حجمها الكبير، لديها سقف أداء محدود عند التقييم النظيف، ويرجع ذلك إلى طبيعة الخصائص المتاحة وغياب بعض المؤشرات السريرية الأكثر تفصيلا.
أوضحت النتائج أن

3.2 بيانات UCI Heart Disease

وLong Beach VA كمعيار طبي معروف في أبحاث أمراض القلب. هذه البيانات أصغر حجما لكنها أكثر ارتباطا بالأدبيات البحثية الطبية الخاصة بتوقع أمراض القلب.
تم استخدام ملفات Cleveland وHungarian وSwitzerland

3.3 بيانات OpenML Heart Disease Comprehensive 1190

وHungarian وSwitzerland وLong Beach VA وStatlog. تحتوي على 1190 سجلا وتوفر توازنا جيدا بين الحجم والقيمة السريرية للخصائص.
تم استخدام هذه البيانات كأقوى معيار نهائي لأنها تجمع مصادر متعددة: Cleveland

4. معالجة البيانات والهندسة المبدئية للخصائص

تم تنفيذ خطوات تنظيف وتجهيز مختلفة حسب كل مجموعة بيانات:

- إزالة القيم غير المنطقية في ضغط الدم والطول والوزن والعمر في بيانات Kaggle.
- حساب مؤشرات مشتقة مثل BMI وpulse pressure وmean arterial pressure.
- تحويل الهدف إلى تصنيف ثنائي في بيانات UCI.
- التعامل مع القيم الناقصة باستخدام median للخصائص الرقمية وmost frequent للخصائص الفئوية.
- استخدام StandardScaler للخصائص الرقمية وOneHotEncoder للخصائص الفئوية.
- الحفاظ على فصل واضح بين طبقة البيانات، طبقة التجهيز، طبقة النمذجة، وطبقة التقييم.

5. النماذج التي تم اختبارها

شملت التجارب النماذج التالية:

- Logistic Regression
- Naive Bayes
- Support Vector Machines
- Random Forest
- Extra Trees
- Gradient Boosting
- AdaBoost
- XGBoost
- LightGBM
- CatBoost
- Multi-layer Perceptron
- Soft Voting Hybrid
- Stacking Hybrid
- Feature Selection + XGBoost/CatBoost
- Feature Tokenizer Transformer

6. النماذج الهجينة

تم بناء نموذجين هجينين أساسيين:

6.1 Soft Voting Hybrid

احتمالات التنبؤ من عدة نماذج قوية. الفكرة أن كل نموذج قد يتعلم نمطا مختلفا من البيانات، وبالتالي فإن دمج الاحتمالات يساعد على تحسين الثبات ورفع ROC-AUC.
يعتمد هذا النموذج على دمج

6.2 Stacking Hybrid

على مجموعة نماذج أساسية، ثم يستخدم نموذجا نهائيا يتعلم من مخرجات هذه النماذج. الهدف هو الاستفادة من نقاط قوة كل نموذج بدل الاعتماد على مصنف واحد فقط. يعتمد هذا النموذج

7. النتائج عبر مجموعات البيانات

dataset	model	accuracy	f1	recall	roc_auc
Kaggle CVD 70k	GradientBoostingClassifier	72.23%	73.93%	79.58%	80.13%
UCI Cleveland	Random Forest	83.49%	81.42%	79.10%	91.37%
UCI All Processed	SVC RBF	82.03%	83.64%	83.07%	87.95%
OpenML Heart 1190	Tuned CatBoostClassifier	94.26%	94.56%	94.86%	96.78%
OpenML Heart 1190	Tuned Soft Voting Hybrid	94.17%	94.52%	95.23%	97.21%
OpenML Heart 1190 + Feature Selection	Mutual Information SelectKBest + Tuned XGB	90.56%	91.00%	90.24%	95.14%
OpenML Heart 1190 + FT-Transformer	Feature Tokenizer Transformer	88.24%	89.23%	92.06%	92.80%

Kaggle حققت أداء أقل نسبيا، بينما قدمت بيانات OpenML Heart 1190 أفضل نتائج عامة. هذا يدعم قرار توسيع الدراسة بدل الاكتفاء بمجموعة بيانات واحدة. توضح النتائج أن بيانات

8. نتائج النماذج المحسنة والهجينة

model	accuracy	accuracy_std	precision	recall	f1	roc_auc	roc_auc_std	pr_auc
Tuned CatBoostClassifier	94.26%	2.08%	94.38%	94.86%	94.56%	96.78%	1.55%	95.86%
Tuned Soft Voting Hybrid	94.17%	1.59%	93.89%	95.23%	94.52%	97.21%	1.43%	96.63%
Tuned Stacking Hybrid	94.06%	1.68%	94.03%	94.86%	94.40%	97.19%	1.44%	96.61%
Tuned LGBMClassifier	93.81%	1.72%	93.90%	94.49%	94.15%	96.90%	1.60%	96.00%
Tuned ExtraTreesClassifier	93.42%	1.55%	93.13%	94.60%	93.81%	96.92%	1.05%	96.50%
Tuned XGBClassifier	93.28%	1.76%	93.28%	94.12%	93.66%	96.37%	1.63%	95.41%
Tuned Random Forest	93.05%	1.83%	92.96%	94.06%	93.47%	96.75%	1.45%	96.08%

بينما حقق Soft Voting Hybrid أعلى ROC-AUC. لذلك يمكن عرض CatBoost كأقوى نموذج منفرد، وSoft Voting/Stacking كأقوى إطار هجين. حقق CatBoost أعلى دقة تقريبا،

9. اختبار Holdout النهائي

model	accuracy	precision	recall	f1	roc_auc	pr_auc
Tuned Stacking Hybrid	92.86%	95.04%	91.27%	93.12%	97.40%	96.90%
Tuned Soft Voting Hybrid	92.86%	94.31%	92.06%	93.17%	97.36%	96.91%
Tuned CatBoostClassifier	92.44%	95.00%	90.48%	92.68%	96.75%	96.08%

يوضح اختبار holdout أن النماذج الهجينة احتفظت بأداء قوي خارج التحقق المتقاطع، خاصة من حيث ROC-AUC وPR-AUC.

10. تحليل اختيار الخصائص

model	accuracy	accuracy_std	precision	recall	f1	roc_auc	roc_auc_std	pr_auc
Mutual Information SelectKBest + Tuned XGB	90.56%	2.40%	91.87%	90.24%	91.00%	95.14%	2.02%	94.22%
ANOVA SelectKBest + Tuned CatBoost	88.49%	2.76%	89.71%	88.49%	89.00%	93.74%	2.09%	93.38%
ANOVA SelectKBest + Tuned XGB	87.42%	2.96%	89.12%	86.96%	87.94%	92.71%	2.61%	92.72%

الخصائص مع النماذج المحسنة والهجينة. هذه نتيجة مهمة لأنها توضح أن الخصائص الكاملة في OpenML تحمل معلومات مفيدة، وأن حذف جزء منها قد يقلل الأداء. ANOVA وMutual Information SelectKBest مع XGBoost وCatBoost. أظهرت النتائج أن تقليل عدد الخصائص لم يتفوق على استخدام جميع تم اختبار SelectKBest

11. Feature Tokenizer Transformer

| model | epochs | best_valid_auc | accuracy | precision | recall | f1 | roc_auc | pr_auc |

| Feature Tokenizer Transformer | 24 | 94.07% | 88.24% | 86.57% | 92.06% | 89.23% | 92.80% | 90.33% |

كنموذج تعلم عميق للبيانات الجدولية. يقوم النموذج بتحويل كل خاصية طبية إلى token embedding ثم يستخدم self-attention لتعلم العلاقات بين الخصائص. تمت إضافة Feature Tokenizer Transformer

في البيانات الطبية الجدولية الصغيرة أو المتوسطة، حيث تكون CatBoost وXGBoost وLightGBM غالبا أكثر كفاءة في التعلم من عدد محدود من السجلات. حقق النموذج أداء جيدا كخط أساس عميق، لكنه لم يتفوق على نماذج boosting والهجين. وهذا متوقع

12. تفسير النموذج باستخدام SHAP

| feature | mean_abs_shap |

| chest_pain_type_4.0 | 1.0901115744116048 |

| ST_slope_2.0 | 0.9824976081888284 |

| oldpeak | 0.8601912077693442 |

| ST_slope_1.0 | 0.8558491137864163 |

| cholesterol | 0.7732620850698556 |

| max_heart_rate | 0.5968171613736415 |

| exercise_angina_1.0 | 0.5889404130214758 |

| sex_1.0 | 0.5165384956746343 |

| age | 0.494082307021947 |

| resting_bp_s | 0.4269255127930438 |

| sex_0.0 | 0.354253630880887 |

| resting_ecg_2.0 | 0.3233158089041967 |

exercise وsex وage وresting blood pressure. هذه النتائج مهمة لأنها تربط أداء النموذج بعوامل طبية يمكن تفسيرها بدل الاكتفاء برقم الدقة فقط. تحليل SHAP أن أهم العوامل المؤثرة في قرارات النموذج تشمل chest pain type وST slope وoldpeak وcholesterol وmax heart rate وangina أظهر

13. المقارنة مع الأعمال المنشورة

| study | dataset | method | accuracy | sensitivity_recall | roc_auc | comparison_type | source |

| Ensemble learning with explainable AI, Scientific Reports 2025 | HDDC / D1 | Stacking ensemble | 91% | Not reported in summary line | 0.92 | Same family of heart-disease benchmark; protocol differs | https://www.nature.com/articles/s41598-025-97547-6 |

| Ensemble learning with explainable AI, Scientific Reports 2025 | UHDD / D2 | Stacking ensemble | 98% | Not reported in summary line | 0.97 | Same family of heart-disease benchmark; protocol differs | https://www.nature.com/articles/s41598-025-97547-6 |

| Optimized Ensemble Learning with XAI, Information 2024 | Cleveland | Bayesian-optimized XGBoost + SHAP | 98.40% | 98.90% | Not reported in Table 5 | Dataset overlaps with UCI Cleveland; protocol differs | https://pdfs.semanticscholar.org/3747/6c8130d1b63c9b1894870c9243026c98a24e.pdf |

| Optimized Ensemble Learning with XAI, Information 2024 | Framingham | Bayesian-optimized XGBoost + SHAP | 95.90% | 97.50% | Not reported in Table 6 | Different dataset; indirect comparison only | https://pdfs.semanticscholar.org/3747/6c8130d1b63c9b1894870c9243026c98a24e.pdf |

| This work | OpenML Heart 1190 | Tuned CatBoostClassifier | 94.26% | 94.86% | 0.9678 | Our reproducible repeated-CV result | Generated by this notebook |

| This work | OpenML Heart 1190 | Tuned Soft Voting Hybrid | 94.17% | 95.23% | 0.9721 | Our reproducible repeated-CV hybrid result | Generated by this notebook |

| This work | OpenML Heart 1190 + FT-Transformer | Feature Tokenizer Transformer | 88.24% | 92.06% | 0.9280 | Our deep tabular baseline | Generated by this notebook |

الأعمال قد تعتمد على مجموعة بيانات صغيرة، أو split واحد، أو تحسين مكثف لنموذج معين مثل XGBoost، أو اختيار خصائص قد يرفع النتيجة على بيانات محددة. بعض الأعمال المنشورة تعرض أرقاما مرتفعة جدا مثل 98% على Cleveland أو UHDD. الفرق الأساسي أن هذه

مجموعة البيانات، ونفس طريقة التقسيم، ونفس خطوات التجهيز، ونفس بروتوكول التقييم. لذلك تم فصل نتائج الأعمال المنشورة عن النتائج التي تم توليدها من النوتبوك. المقارنة المباشرة لا تكون صحيحة إلا عند استخدام نفس

في هذا العمل تم التركيز على إطار أوسع وأكثر قابلية للتكرار:

- استخدام أكثر من مجموعة بيانات.

- مقارنة مصادر بيانات مختلفة.

- استخدام repeated stratified cross-validation بدل الاعتماد فقط على split واحد.

- اختبار نماذج منفردة وهجينة.

- إضافة تفسير SHAP.

- إضافة Feature Tokenizer Transformer.

- توضيح حدود كل Dataset بدل الاكتفاء بأفضل رقم فقط.

لذلك فإن قوة العمل ليست فقط في الوصول إلى رقم عال، بل في بناء إطار بحثي شامل يوضح متى ولماذا تتحسن النتائج، وما حدود كل مجموعة بيانات.

14. الإسهام البحثي

```
| issue | project_response | evidence |

| Using multiple datasets alone is not a sufficient contribution. | Treat multiple datasets as validation evidence, not as the main novelty. |
Kaggle, UCI, OpenML, feature-selection, and FT-Transformer results are reported separately. |

| A clear technical contribution is required. | Propose a unified explainable hybrid framework. | Soft Voting and Stacking hybrids combine tuned
boosting and tree-ensemble learners. |

| The model should not be a black box. | Add SHAP explainability for the final tree-based model. | SHAP identifies chest pain type, ST slope,
oldpeak, cholesterol, max heart rate, and exercise angina. |

| Deep learning should be represented if the thesis claims modern AI. | Add Feature Tokenizer Transformer for tabular clinical data. |
FT-Transformer is evaluated as a deep tabular baseline and compared with boosting/hybrid models. |

| Feature engineering/selection should be evaluated, not assumed. | Add ANOVA and mutual-information SelectKBest experiments. | Feature selection
is shown to reduce dimensionality but does not outperform the full hybrid framework. |

| High published accuracies may come from different protocols. | Separate published results from reproducible notebook-generated results. |
Published-work comparison table includes a comparison-type note for each row. |
```

يمكن تلخيص الإسهام البحثي في النقاط التالية:

1. بناء إطار تجريبي متعدد البيانات للتنبؤ بأمراض القلب.
2. تحليل حدود بيانات Kaggle وإظهار أن انخفاض الأداء مرتبط بطبيعة الخصائص وليس فقط باختيار النموذج.
3. استخدام OpenML Heart 1190 كمصدر أقوى يجمع أكثر من benchmark طبي.
4. بناء نماذج هجينة Soft Voting وStacking اعتمادا على نماذج boosting وtree ensembles.
5. إضافة SHAP explainability لتفسير العوامل الطبية المؤثرة.
6. دراسة أثر feature selection على الأداء.
7. إضافة Feature Tokenizer Transformer كخط أساس حديث للتعلم العميق على البيانات الجدولية.
8. اعتماد تقييم أكثر ثباتا من خلال repeated stratified cross-validation واختبار holdout.

هو الإطار الموحد القابل للتفسير الذي يجمع بين النماذج الهجينة، وSHAP، وFeature Selection، وFeature Tokenizer Transformer، والتقييم القوي. استخدام أكثر من Dataset لا يتم تقديمه كإسهام بحثي مستقل، ولكنه دليل تحقق validation على ثبات الإطار المقترح. الإسهام الأساسي

15. الخلاصة

بينما حقق Soft Voting Hybrid أعلى ROC-AUC. كما أوضح تحليل SHAP أن قرارات النموذج تعتمد على عوامل طبية منطقية، مما يزيد قابلية تفسير النتائج. أظهرت النتائج أن أفضل أداء تحقق على بيانات OpenML Heart Disease Comprehensive 1190. حقق نموذج CatBoost المحسن أعلى دقة تقريبا،

وإن لم يتفوق على boosting models، لأنها تثبت أن الدراسة لا تقتصر على النماذج التقليدية فقط، بل تقارن أيضا اتجاهات حديثة في deep tabular learning. إضافة Feature Tokenizer Transformer تمثل امتدادا حديثا للعمل، حتى

بناء على ذلك، أصبح العمل إطارا بحثيا متكاملا يجمع بين الأداء، المقارنة بين البيانات، النماذج الهجينة، قابلية التفسير، وتحليل الخصائص.