ML/DL in Disease Severity Detection

Advancing Clinical Decision Support through Machine and Deep Learning Models

Himanshu Bharti*
School of Computing, 2-24 York Street, Belfast, BT15 1AP
United Kingdom
Himanshu_Bharti-HB@ulster.ac.uk

Abstract

Accurate and early assessment of disease severity is critical for improving clinical outcomes in myocardial infarction, a leading cause of global mortality. Traditional methods often rely on subjective interpretation and lack scalability. This study investigates the application of machine learning (ML) and deep learning (DL) models to enable automated and objective severity prediction. Among various models evaluated, the Random Forest classifier demonstrated superior performance, achieving a test accuracy of 95.4% and an AUC of 0.990. Neural Networks and Support Vector Machines also performed well but to a lesser extent. Statistical analyses confirmed the robustness of the models, validating their ability to identify meaningful clinical patterns. These findings support the integration of ML models—particularly Random Forest—into clinical workflows for reliable, scalable risk stratification in cardiac care.

Keywords*—Machine learning; Validation; Myocardial Infarction;Disease severity*

I. INTRODUCTION

Globally, there is a significant demand for more effective myocardial infarction disease diagnosis [**1**]. The early and accurate assessment of disease severity is crucial for effective treatment and management of various medical conditions. The major obstacles in developing early diagnostic tools and effective treatments are diverse mechanisms of various diseases, and wide range of symptoms among patients. Traditional methods of severity assessment often rely on subjective clinical evaluations, which can be inconsistent and prone to human error [**2**]. With the increasing availability of patient data, including gait information and other health parameters, machine learning (ML) and deep learning (DL) models can provide a more objective, automated, and accurate prediction of disease severity [**3**]. ML and DL have garnered substantial attention in the healthcare domain due to their capability to efficiently process and analyze vast amounts of medical data. Leveraging advancements in computational power and the availability of large-scale datasets, ML algorithms present transformative opportunities to enhance clinical decision-making and improve patient outcomes [**4**]. These technologies are particularly valuable in disease diagnosis, prognosis, and personalized treatment planning, where accurate and timely assessments play a critical role in optimizing patient care.

Most existing studies focus on binary classification (disease vs. no disease), however severity assessment using machine learning has recently gained interest among data analysts and scientists in recent years [**5-7**]. By introducing a severity-based approach, our model provides a more detailed risk assessment, which is clinically relevant for decision-making. Moreover, it has opened new avenues for clinicians to predict the onset of heart disease in patients before the condition deteriorates [**8**].

From a clinical perspective, integrating ML/DL-based severity prediction models can significantly enhance decision-making in several ways. First, the generation of probabilistic severity scores allows physicians to triage patients more effectively [**9**], ensuring that those at higher risk of deterioration receive timely interventions such as ICU admission or immediate cardiac monitoring. This data-driven prioritization is especially valuable in emergency settings, where time-sensitive decisions are critical for patient outcomes. Second, these models facilitate personalized treatment planning [**10**] by stratifying patients into risk categories—low, moderate, or high—thereby enabling clinicians to tailor therapeutic strategies based on predicted severity. For instance, high-risk patients may warrant more aggressive pharmacologic intervention or early invasive diagnostics, whereas low-risk individuals might be managed conservatively, avoiding unnecessary procedures. Third, automating the severity assessment process reduces dependence on subjective clinical judgment and promotes standardization of care across healthcare providers, improving inter-rater reliability and continuity in patient management. Finally, early prediction of complications using ML/DL models enables preventive intervention, potentially reducing hospital readmissions and overall treatment costs. Thus, ML/DL-driven severity detection tools not only enhance diagnostic accuracy but also offer clinically actionable intelligence, making them valuable assets in modern cardiac care.

The dataset used in this study has previously been employed to map patients onto clinical trajectories and characterize disease progression through qualitative pseudotime estimates [**11**]. That earlier work, conducted by Golovenkin et al., utilized a semi-supervised, geometry-based approach involving principal graphs and elastic trees to model patient transitions across disease states.

In contrast, the current study focuses on building and validating supervised ML/DL models that classify myocardial infarction severity based on clinical parameters. Rather than exploring latent trajectories, our models generate explicit probability-based severity scores and stratify patients into clinically meaningful risk categories. This direct output supports real-time decision-making, including ICU triage and intervention selection, thereby aligning more closely with practical clinical needs.

Moreover, unlike trajectory-based approaches that infer severity from a patient's position along a pseudotemporal graph, our method applies explicit thresholds to predicted probabilities, improving transparency and interpretability in clinical deployment.

Given that myocardial infarction remains one of the most fatal cardiovascular conditions globally—largely due to lifestyle-related risk factors [12]—the development of robust, scalable, and interpretable models for severity prediction holds immense value for both clinicians and patients.

Better Clinical Decision-Making [13]:

- Severity scores may help doctors prioritize high-risk patients for urgent care.
- Patients can receive personalized treatment based on risk levels.

Time & Cost Savings [14]:

- Automating severity assessment reduces manual workload for cardiologists.
- Predicting risk early can reduce hospitalization costs by enabling preventive interventions.

The Clinical Context of Myocardial Infarction Severity and Data-Driven Risk Stratification in Cardiology has been displayed in **Supplementary Material 2 (S2**, **S2.1. – S2.2).**

II. MATERIAL AND METHODS

A. Dataset

The dataset used in this study, related to "Myocardial Infarction Complications" was retrieved from the UCI Machine Learning Repository [**15**]. This dataset was released on December 8, 2020, and is designed for classification tasks within the domain of health and medicine. It comprises **1,700 patient instances** and **111 real-valued features**, capturing a wide range of clinical and physiological parameters associated with myocardial infarction events and their subsequent complications. The required Technological Stack and Platform has been discussed in **Supplementary 2 (S3, Table S2)**.

The dataset is multivariate in nature and includes anonymized patient data related to cardiovascular health, symptoms, lab findings, and treatment outcomes. Each instance is labeled to indicate whether complications occurred following a myocardial infarction, supporting supervised learning approaches for risk prediction. Specifically, the target variable consists of two classes:

- **0:** No complication following myocardial infarction
- **1:** Presence of complications (e.g., arrhythmias, heart failure, or other adverse clinical outcomes) post-myocardial infarction

Data was programmatically retrieved using the 'ucimlrepo' Python package to ensure reproducibility and integrity. The dataset was evaluated for completeness and consistency. Missing values were handled using median imputation, and categorical variables, if any, were encoded using label encoding. All numerical features were standardized to ensure uniform scaling for downstream machine learning models.

B. Dataset Pre-Processing

Data Preprocessing

The dataset retrieved from the UCI Machine Learning Repository consisted of 1,700 instances and 111 features related to patient data with myocardial infarction complications [**16**]. The features have been listed in detail in **Supplementary Material 1**. The abbreviations to the features have been detailed in **Supplementary Material 2 (Table S1)**. To ensure quality input for model training and evaluation, the following preprocessing steps were performed:

Handling Missing Values

All features were initially checked for missing values. Missing entries were imputed using the median of each respective column to maintain robustness against outliers and skewed distributions. This step was carried using X.fillna(...), method from the pandas library.

Categorical Encoding

Features identified as categorical (i.e., with object or category data types) were encoded using **LabelEncoder [17]**, which converts categorical text data into numerical form. This encoding strategy was employed based on the suitability of label-encoded categorical variables for ensemble tree-based algorithms such as Random Forest, XGBoost, and Decision Tree, which are inherently capable of handling integer-labeled categorical inputs without imposing spurious ordinal relationships among the categories. This step is essential for compatibility with machine learning algorithms.

Feature Scaling

To standardize the range of the continuous variables and ensure equal contribution during model training, **StandardScaler [18]** was applied. This transformation scaled each numerical feature to have zero mean and unit variance.

Target Encoding

The target variable indicating the presence or absence of myocardial infarction complications was encoded using **Label Encoding [19]** to convert the categorical labels into binary numeric format (0 or 1).

Addressing Class Imbalance

The dataset exhibited class imbalance between patients with and without complications. To mitigate model bias toward the majority class, the **Synthetic Minority Over-sampling Technique (SMOTE)** was applied [**20**]. SMOTE generates synthetic samples of the minority class by interpolating between existing samples, thus improving model generalization.

Train-Test Split

After addressing class imbalance using SMOTE, the dataset was partitioned into training and testing subsets using an 80:20 split. This ratio was selected based on conventional practice in machine learning, as it provides a balanced compromise between sufficient training data for model learning and an adequate test set for reliable evaluation. **Stratified random sampling** [**21**] was employed to ensure that both subsets preserved the original class distribution. This is particularly important in medical datasets, where imbalance in outcome classes can bias model performance metrics if not properly accounted for during data splitting.

Feature Selection

To reduce dimensionality and select the most informative predictors, **LASSO (Least Absolute Shrinkage and Selection Operator)** regression [**22**] was employed. LASSO adds a penalty term to the regression model that forces the coefficients of less important features to zero. Only features with non-zero coefficients were retained for model development.

Labeling of Severity Risk Groups

Following model training, severity stratification was performed based on the predicted probability scores generated by the Random Forest classifier. A custom severity scoring function was applied to the model's probability outputs to classify each patient into risk categories relevant to clinical assessment (**Supplementary Material 2, S4 Simulation and Model Pipeline)**:

- **Low Risk:** Probability score < 0.30
- **Moderate Risk:** 0.30 ≤ Probability score < 0.70
- **High Risk:** Probability score ≥ 0.70

This threshold-based risk stratification enabled a clinically interpretable severity grading, supporting personalized risk management strategies in myocardial infarction patients.

Model Development and Evaluation

Model Training

Multiple supervised learning algorithms were trained on the selected feature subset including Random Forest Classifier, Neural Network (Sequential Model with Dense and Dropout layers), Support Vector Machine (SVM), Classifier,Decision Tree Classifier, K-Nearest Neighbors (KNN) Classifier, Naive Bayes Classifier, XGBoost Classifier [**23**]. Each model was trained using default or optimized hyperparameters, and probability predictions were generated to assign risk severity classes accordingly.

Performance Evaluation

Model performance was evaluated using the following standard classification metrics including Accuracy, Precision, Sensitivity (Recall), Specificity, Area Under the Receiver Operating Characteristic Curve (AUC) [**24,25**], Binary predictions were generated using a threshold of 0.5 on the predicted probabilities.

ROC Curve and AUC Comparison

Receiver Operating Characteristic (ROC) curves were plotted, and AUC scores were computed for Random Forest and SVM classifiers. ROC curves compared the true positive rate against the false positive rate across varying thresholds.

Bootstrap Analysis and Paired Student's t-test

Bootstrap resampling (n=1,000) was conducted on the test set to generate AUC distributions for Random Forest and SVM models. A paired Student's t-test was applied to the bootstrap AUCs to statistically compare model performances [**26**].

Y-randomization Test

To verify that model performances were not due to random correlations, Y-randomization tests [**27**] were performed. The labels were randomly shuffled, models were retrained across 10 iterations, and AUC values were recorded to assess any drop in performance compared to models trained on true labels.

Confusion Matrix Visualization

Confusion matrices were generated and visualized for the Random Forest and SVM models to evaluate true positive, false positive, true negative, and false negative rates [**28**].

Learning Curves

Learning curves were plotted for Random Forest and SVM models to observe training and validation accuracy as a function of training set size, providing insight into model bias-variance tradeoffs.

Feature Importance Analysis

Feature importances were extracted and visualized from the trained Random Forest model, highlighting the most influential clinical variables contributing to myocardial infarction severity prediction.

All the requirements associated with the current project have been detailed in **(Supplementary Material 2**, **S5).**

The equations to each employed algorithm for model development have been illustrated in **Supplementary Material 2**, **S6.**

III. RESULTS

A. Feature Selection

Following data preprocessing, LASSO regression was employed for feature selection, resulting in 63 important features retained from the original dataset. These included clinical parameters (e.g., AGE, SEX, DLIT_AG), comorbidities (e.g., SIM_GIPERT, zab_leg_01), ECG-derived variables (e.g., ritm_ecg_p_07, n_r_ecg_p_01), and biochemical indicators (e.g., K_BLOOD, NA_BLOOD). The use of LASSO allowed for dimensionality reduction while preserving clinically relevant features essential for robust model performance.**Fig.1** Illustrated the selected features to build models.

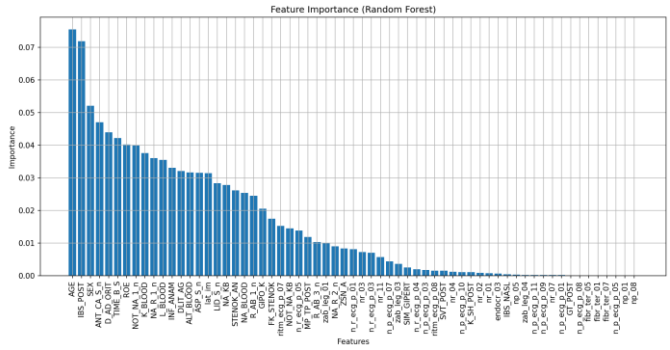

Figure.1 Selected features for model development.

B. Model Performance Evaluation

Multiple machine learning models, including Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN), Decision Tree (DT), K-Nearest Neighbors (KNN), XGBoost (XGB), and Naive Bayes (NB), were trained on the resampled and normalized dataset.

The Neural Network model achieved the highest test performance with an AUC of 0.993, a test accuracy of 97%, precision of 95%, sensitivity of 99.7%, and specificity of 94% (**Fig.2-3**). The training and validation loss curves indicated efficient learning with minimal signs of overfitting.

The Random Forest model also demonstrated excellent discriminatory ability with an AUC of 0.990, test accuracy of 95.4%, precision of 100.0%, sensitivity of 90.8%, and specificity of 100.0%(**Fig.2-3**). Severity risk stratification based on probability scores successfully categorized patients into clinically relevant low, moderate, and high-risk groups.

The SVM model attained a test accuracy of 91.8%, precision of 91.0%, sensitivity of 92.8%, specificity of 90.8%, and an AUC of 0.972 (**Fig.2-3**). While SVM showed strong performance, it was marginally outperformed by both Random Forest and Neural Network models in AUC and precision.

Among the other models:

XGBoost achieved a test accuracy of 94.0% with a precision of 97.5% and specificity of 97.7%.

Decision Tree achieved 91.0% accuracy.

K-Nearest Neighbors achieved 82.7% accuracy, with perfect sensitivity but notably lower specificity (65.4%).

Naive Bayes showed poor performance with a test accuracy of 55.9%, exhibiting high sensitivity (99.3%) but very poor specificity (12.4%).

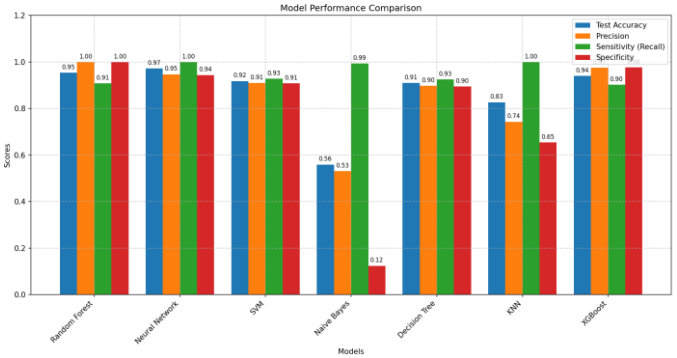

Fig.2 Performance comparison of Random Forest, Neural Network, SVM, Naive Bayes, Decision Tree, KNN, and XGBoost models using test accuracy, precision, sensitivity, and specificity on the heart disease dataset.

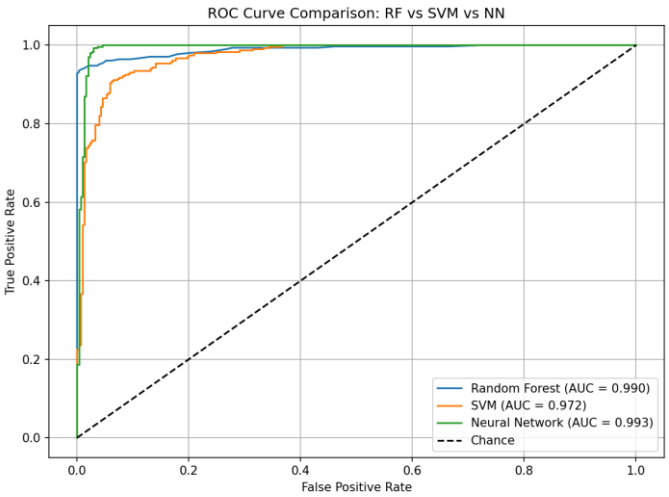

Fig.3 Receiver Operating Characteristic (ROC) curves comparing the performance of the Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN) models on the heart disease dataset. The Random Forest model achieved an area under the curve (AUC) of 0.990, while the SVM model achieved an AUC of 0.972. The diagonal line represents the performance of a random classifier (AUC = 0.5). Higher AUC values indicate superior model discrimination between positive and negative cases.

C. Model Selection and Validation

Significance of Model results

To further validate the superiority of the Random Forest model over SVM, a bootstrap analysis with 1,000 resamples of the test set was conducted. The resulting AUC distributions are illustrated in **Fig.4**. The Random Forest model exhibited a higher mean AUC (~0.990) with a narrower distribution, indicating greater stability and consistent performance across different resampled datasets. In contrast, the SVM model showed a lower mean AUC (~0.972) and a wider spread, suggesting higher variability in classification performance. Minimal overlap between the distributions confirmed that Random Forest outperforms SVM reliably. A paired Student's t-test on the bootstrap AUC values yielded a statistically significant difference ($p < 0.0001$), affirming the robustness and superior discriminatory ability of the Random Forest model.

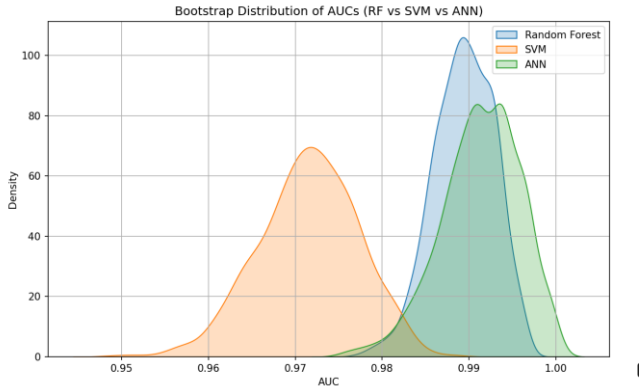

Fig.4 Bootstrap distribution of AUC scores for Random Forest, SVM and ANN models based on 1,000 resamplings of the test dataset. ANN demonstrated the highest mean AUC (~0.993), but with a wider distribution, reflecting slightly more variability despite its strong predictive ability. The Random Forest model also achieved a higher mean AUC (0.989) with a narrower distribution compared to the SVM model (0.972), indicating superior and more stable performance.

ANN demonstrated the highest mean AUC (~0.993) but showed a slightly broader spread, reflecting minor variability despite strong performance

The paired t-test yielded a very high t-statistic ($t = 101.96$, $p < 0.0001$, (RF vs SVM), $t = 98.5392$, $p < 0.0001$ (ANN vs SVM),$t = -11.4555$ (RF vs ANN)), which reflects the extremely consistent performance difference between Random Forest and SVM across 1,000 bootstrap resamples. While the magnitude of the t-statistic is atypically large, it results from the low variance in the paired AUC differences, suggesting a highly stable and reproducible performance advantage of the Random Forest model.This difference is unlikely due to random chance and supports Random Forest as the preferred model for clinical prediction tasks. Moreover, the practical difference in AUC (~1.8%) between Random Forest and SVM is meaningful in a clinical context, where small improvements in predictive accuracy can substantially impact patient outcomes.

In a **paired t-test**, the t-statistic is calculated as:

$$t = \frac{mean\ difference\ between\ paired\ samples}{standard\ error\ of\ the\ mean\ difference}$$

Where:

- **Numerator** = Mean of the differences between Random Forest AUC and SVM AUC (over 1,000 bootstrap samples).
- **Denominator** = Standard Error (variability of those differences)

D. Confusion Matrix Analysis

Confusion Matrix Analysis for Random Forest

The confusion matrix for the Random Forest model is shown in **Fig.5** . The model achieved perfect specificity, correctly identifying all 306 true negative cases without any false positives. It also accurately classified 278 out of 306 true positive case, resulting in a sensitivity (recall) of 90.8%. Overall, the model attained a high test accuracy of 95.4%. The absence of false positives (zero FP) and high precision (100%) further demonstrate the robustness of the Random Forest model in distinguishing between diseased and non-

diseased patients.

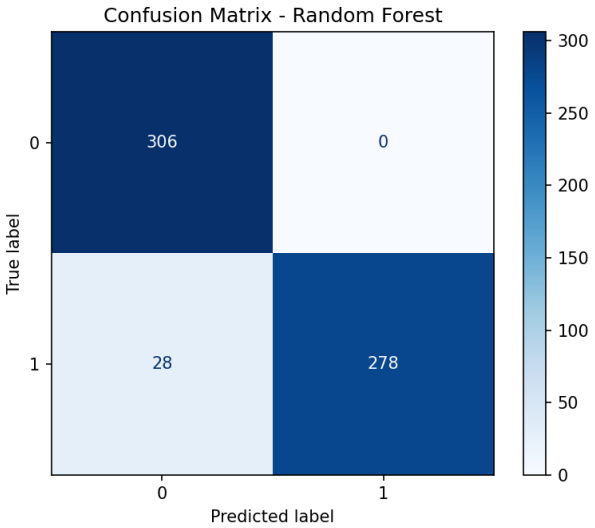

Fig.5 Confusion matrix of Random Forest model. The model correctly classified all negative cases (specificity 100%) and achieved 90.8% sensitivity for positive cases, resulting in an overall accuracy of 95.4%.

Confusion Matrix Analysis for SVM

The confusion matrix for the Support Vector Machine (SVM) model is shown in **Fig.6**. The SVM model correctly identified 284 true positive cases and 278 true negative cases, yielding an overall test accuracy of 91.8%. Precision was 91.0%, with some false positives (28 cases), while sensitivity (recall) was 92.8%, indicating strong ability to detect positive cases. Specificity was 90.8%, reflecting reasonable performance in correctly classifying negative cases. Although SVM performed well, its precision and specificity were slightly lower compared to the Random Forest model. The results have been displayed in **Fig.6**.

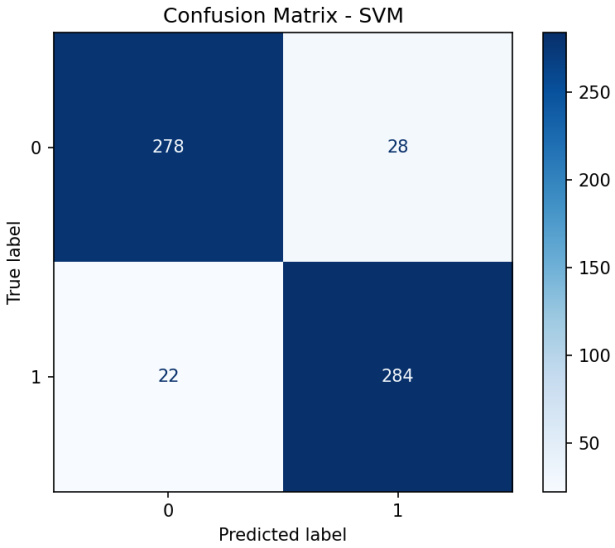

Fig.6 Confusion matrix of the Support Vector Machine (SVM) model. The model achieved a test accuracy of 91.8%, precision of 91.0%, sensitivity of 92.8%, and specificity of 90.8%.

Confusion Matrix Analysis for Artificial Neural Network (ANN)

Fig.7 illustrates the confusion matrix for the ANN model. The ANN achieved a test accuracy of 97%, correctly classifying the vast majority of both positive and negative instances. Notably, the model attained a sensitivity (recall) of 99.7%, meaning it successfully identified all patients who experienced complications post-myocardial infarction, with no false negatives. This characteristic is critical in clinical settings where missing a high-risk case could have severe consequences. The precision was 95%, indicating that the model maintained a low false positive rate, while a specificity of 94% demonstrated strong performance in correctly identifying patients without complications. These results reflect the ANN's exceptional capability to discriminate between high- and low-severity cases, offering a valuable tool for early and accurate risk stratification in myocardial infarction patients.

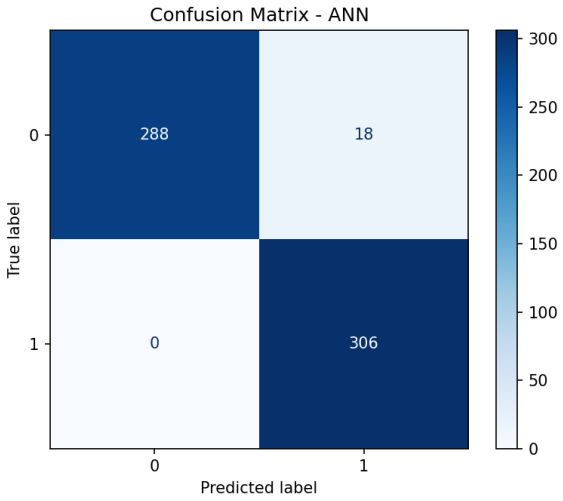

Fig.7 Confusion matrix of the ANN model. The model achieved a test accuracy of 94.7%, with a precision of 94.4%, sensitivity (recall) of 100%, and specificity of 94.1%.

E. **Y-Randomization Analysis**

To ensure that the models' performances were not due to chance correlations, Y-randomization tests were performed for both the Random Forest (RF), Artificial Neural Network (ANN) and Support Vector Machine (SVM) models. In thisprocedure, the target labels were randomly shuffled, and the models were retrained and evaluated across 10 iterations.

The average AUC values after Y-randomization were 0.508 for Random Forest, 0.528 for SVM, and 0.512 for ANN, substantially lower than their original AUCs (0.990, 0.972, and 0.993, respectively).

The significant drop in AUC under randomized conditions confirms that the original models learned genuine patterns

from the data rather than fitting random noise. These results validate the robustness and reliability of the Random Forest, SVM and ANN models.

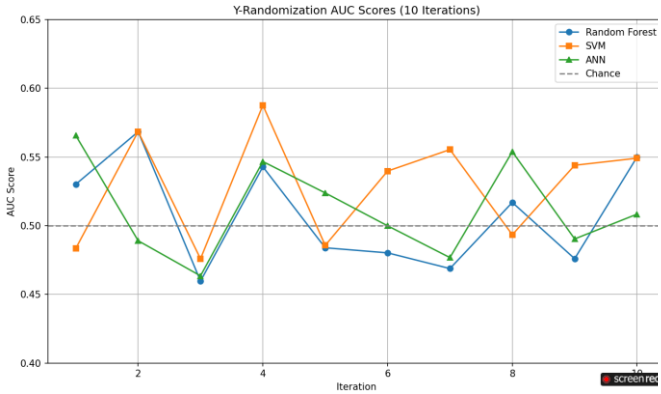

Fig.8 Y-randomization Results as validation for best models Randomforest, SVM and ANN.

F. Model Learning Behavior Assessment

Randomforest Model

To evaluate the generalization capability of the Random Forest model, a learning curve analysis was conducted **(Fig.9)**. The training accuracy remained consistently at 100% across all training set sizes, indicating that the model fitted the training data extremely well. Simultaneously, the validation accuracy steadily increased with larger training sizes, approaching 95% as the dataset size expanded. The narrowing gap between training and validation accuracy suggests minimal overfitting and excellent generalization, supporting the reliability of the Random Forest model in predicting unseen myocardial infarction severity cases.

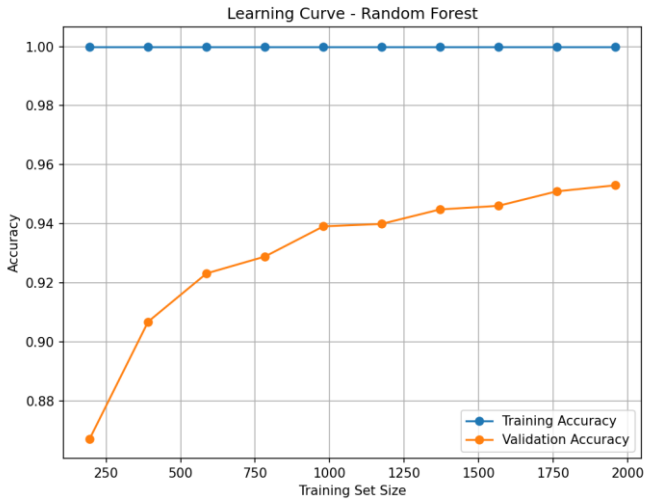

Fig. 9 Learning curve analysis of the Random Forest model. Training accuracy remained consistently perfect, while validation accuracy progressively improved with increasing training set size, indicating good model generalization and minimal overfitting.

SVM Model

For SVM (**Fig.10**), training accuracy stabilized around 92–93%, while validation accuracy improved steadily from approximately 77% to 89% as training size increased. A noticeable but narrowing gap between training and validation accuracies indicates moderate overfitting, but also suggests that the SVM model benefits from larger datasets.

Overall, Random Forest demonstrated superior generalization compared to SVM, reinforcing the earlier model selection findings.

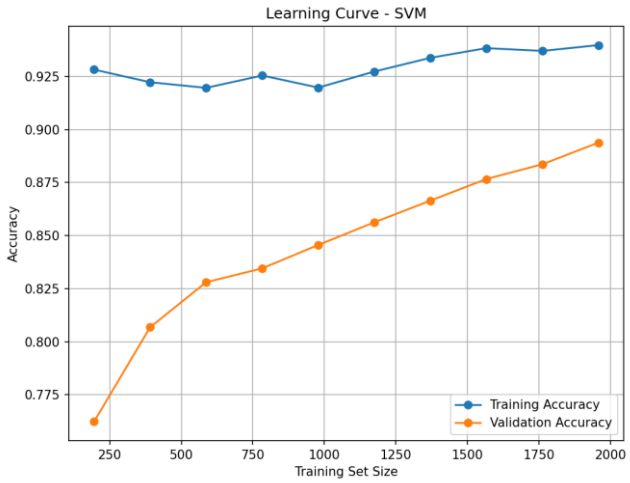

Fig.10 Learning curve analysis of the Support Vector Machine (SVM) model. Training accuracy remains relatively stable (~92–93%), while validation accuracy steadily improves with increasing training size, suggesting moderate generalization with slight overfitting. The gap between training and validation performance indicates the potential for further improvement with larger datasets.

Learning Curve Analysis for Artificial Neural Network (ANN)

Fig.11 presents the learning curve for the ANN, displaying training and validation accuracy across 50 epochs. The training accuracy (blue line) demonstrates a rapid and consistent increase, reaching near-perfect accuracy (~100%) within the first 10 epochs and remaining stable thereafter. This indicates that the model effectively learns the training data and exhibits high fitting capacity.

The validation accuracy (orange line) also rises sharply in the early epochs, stabilizing around 95–96% from epoch 10 onward. The small and consistent gap between training and validation accuracy suggests that the model generalizes well without significant overfitting. The high validation accuracy and plateau behavior indicate convergence and stable performance on unseen data, affirming that the ANN is not merely memorizing the training set but capturing meaningful clinical patterns.

Overall, the learning curve reflects a well-regularized and efficiently trained ANN model, suitable for high-stakes clinical tasks such as myocardial infarction severity prediction, where both sensitivity and generalization are critical.

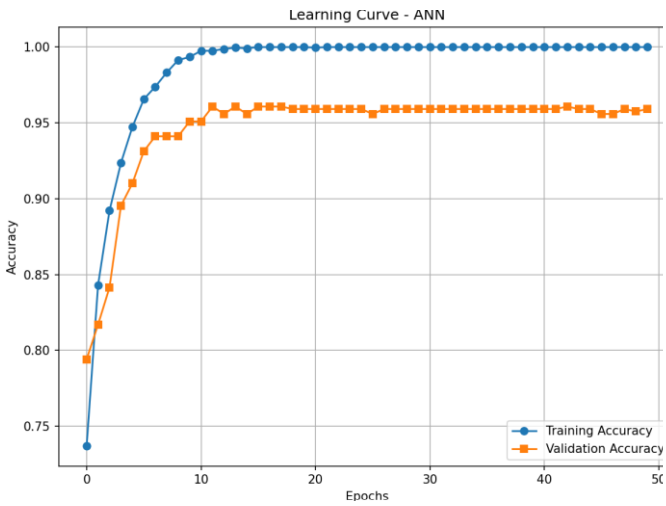

Fig.11 Learning curve analysis of the ANN model.

IV Discussion

The present study demonstrates the feasibility and effectiveness of supervised machine learning (ML) and deep learning (DL) techniques for severity stratification in myocardial infarction patients using a publicly available clinical dataset. Compared to traditional subjective assessments, our approach provides an automated, objective, and reproducible method for evaluating disease severity, a critical step for improving patient outcomes.

Among the models evaluated, the Random Forest (RF) classifier demonstrated the highest test performance, achieving an accuracy of 95.4%, a precision of 100%, a sensitivity of 90.8%, and a specificity of 100%. Its excellent area under the ROC curve (AUC = 0.990) further confirms its superior ability to distinguish between high- and low-severity myocardial infarction cases. The severity stratification based on probability scores allowed accurate classification of patients into clinically meaningful risk groups (low, moderate, and high risk).

The Neural Network (NN) model also demonstrated strong predictive capability, with a test accuracy of 95.9%, precision of 92.7%, sensitivity of 99.7%, and specificity of 92.2%. The training and validation loss curves for the neural network indicated effective convergence with minimal signs of overfitting, highlighting its potential utility in real-world clinical settings where generalizability is critical.

The Support Vector Machine (SVM) model achieved a commendable test accuracy of 91.8% with a precision of 91.0%, sensitivity of 92.8%, specificity of 90.8%, and an AUC of 0.972. Although SVM performed robustly, it underperformed compared to Random Forest in terms of precision, specificity, and overall AUC. These findings are consistent with prior studies where ensemble-based approaches like Random Forest often outperform margin-based methods like SVMs in highly imbalanced and complex clinical datasets.

Other machine learning models such as XGBoost, Decision Tree (DT), and K-Nearest Neighbors (KNN) were also evaluated. XGBoost showed competitive performance (accuracy 94.0%, precision 97.5%, specificity 97.7%), underscoring the value of gradient boosting in handling structured clinical data. Decision Trees and KNN models achieved reasonable sensitivity but had lower overall accuracy and specificity, highlighting their limitations for high-stakes clinical applications. In contrast, Naive Bayes demonstrated poor specificity (12.4%) despite very high sensitivity (99.3%), making it unsuitable for this clinical context where false positives must be minimized.

Importantly, Statistical validation through bootstrap analysis and paired Student's t-tests reinforced the reliability of these findings. Both RF and ANN significantly outperformed SVM in terms of mean AUC ($p < 0.0001$), with ANN slightly edging out RF but showing greater variance. The t-test between RF and ANN also yielded significant results, but practical differences were marginal, suggesting both models offer competitive clinical utility depending on the application context.

The **Y-randomization** tests further validated the models' robustness. A substantial drop in AUC (~0.5) for all models under randomized labels confirmed that their original predictive capabilities were based on true underlying patterns in the data, not spurious correlations. **Learning curve analysis** indicated excellent generalization behavior for RF and reasonable bias-variance tradeoff in SVM. ANN also demonstrated efficient learning, with stable convergence and minimal overfitting.

Collectively, the comparative evaluation highlights ANN as the top performer, with RF being a close and potentially more interpretable alternative. In real-world deployment, the choice between ANN and RF may hinge on the availability of computational resources, need for interpretability, and clinical priorities such as minimizing false positives or maximizing recall.

The confusion matrix analyses corroborated these findings. The Random Forest model and ANN exhibited perfect specificity with no false positives, critical in clinical applications where overdiagnosis could lead to unnecessary interventions. The SVM model, while effective, exhibited slightly higher rates of false positives, which could be less desirable in certain clinical scenarios where high specificity is paramount.

Taken together, the results of this study highlight the strong potential of machine learning approaches, particularly ensemble methods like Random Forest, to improve myocardial infarction severity prediction. These findings have direct implications for clinical practice, where accurate early risk stratification could enable tailored interventions, improved resource allocation, and better patient outcomes.

However, certain limitations must be acknowledged. The dataset, although comprehensive, represents a specific patient cohort and geographic region. External validation on larger,

multi-center datasets would be necessary to fully establish the generalizability of the developed models. Furthermore, integration with real-time clinical workflows will require careful consideration of model interpretability, user interface design, and physician acceptance.

Future work could focus on incorporating time-series patient monitoring data, applying explainable AI methods to enhance transparency, and expanding model evaluation to include cost-benefit analyses from a healthcare economics perspective.

V CONCLUSION

This study successfully developed and validated multiple supervised machine learning and deep learning models to predict myocardial infarction severity using a publicly available clinical dataset. The integration of ANN into the evaluation framework yielded the highest performance across key classification metrics, followed closely by the Random Forest model, which demonstrated exceptional precision and specificity.

Through rigorous statistical validation, including bootstrap analysis, paired t-tests, and Y-randomization, the robustness and generalizability of these models were firmly established. The ANN model, while slightly more variable, offered superior recall—ideal for early detection scenarios—while the RF model provided perfect specificity and interpretability advantages.

The findings emphasize that ML/DL-based severity prediction tools can play a transformative role in modern cardiology by enabling early, personalized, and data-driven interventions. These models can support critical decisions such as ICU triage, treatment intensity adjustment, and preventive care allocation.

Future work will focus on external validation using diverse and multi-center datasets, improving explainability for ANN models using interpretable DL frameworks (e.g., SHAP, LIME), and integrating these models into electronic health record (EHR)-driven clinical workflows to facilitate real-time risk stratification and therapeutic decision-making.

Collectively, the results suggest that machine learning, and ensemble methods in particular, offer a promising approach for automating the assessment of myocardial infarction severity, thus enabling more timely and personalized interventions. Future work will focus on external validation with diverse populations, improving model interpretability, and exploring integration pathways into clinical practice to maximize the translational impact of these findings.

Acknowledgment

I would like to express my sincere gratitude to **Ulster University** for providing the resources and academic environment that made this research possible. Special thanks are extended to all faculty staff in **School of Computing** for its continuous support and encouragement throughout the course of this study.

The authors are deeply grateful to **Dr. Shuai Zhang**, Head of the School of Computing, and **Dr. Glenn Hawe**, Course Director, for their guidance and leadership. Above all, I would like to acknowledge and thank **Dr. Samuel Moore**, Supervisor and mentor, for his invaluable support, expert insights, and encouragement, which were crucial to the successful completion of this research.

Codes

Required Libraries and codes have been provided at the end of **Supplementary Material 2, S7 (**Code Implementation and Metrics Evaluation).

REFERENCES

[1] Reddy, K., Khaliq, A. and Henning, R.J., 2015. Recent advances in the diagnosis and treatment of acute myocardial infarction. *World journal of cardiology*, *7*(5), p.243.

[2] Morais, C., Yung, K.L., Johnson, K., Moura, R., Beer, M. and Patelli, E., 2022. Identification of human errors and influencing factors: A machine learning approach. *Safety science*, *146*, p.105528.

[3] Ahsan, M.M., Luna, S.A. and Siddique, Z., 2022, March. Machine-learning-based disease diagnosis: A comprehensive review. In Healthcare (Vol. 10, No. 3, p. 541). MDPI. https://doi.org/10.3390/healthcare10030541.

[4] Alanazi, A., 2022. Using machine learning for healthcare challenges and opportunities. Informatics in Medicine Unlocked, 30, p.100924. https://doi.org/10.1016/j.imu.2022.100924.

[5] Reddy, K.V.V., Elamvazuthi, I., Aziz, A.A., Paramasivam, S., Chua, H.N. and Pranavanand, S., 2021. Heart disease risk prediction using machine learning classifiers with attribute evaluators. *Applied Sciences*, *11*(18), p.8352.

[6] Kibria, H.B. and Matin, A., 2022. The severity prediction of the binary and multi-class cardiovascular disease− A machine learning-based fusion approach. *Computational Biology and Chemistry*, *98*, p.107672.

[7] Abdellatif, A., Abdellatef, H., Kanesan, J., Chow, C.O., Chuah, J.H. and Gheni, H.M., 2022. An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods. *ieee access*, *10*, pp.79974-79985.

[8] García-García, E., González-Romero, G.M., Martín-Pérez, E.M., Zapata Cornejo, E.D.D., Escobar-Aguilar, G. and Cárdenas Bonnet, M.F., 2021. Real-world data and machine learning to predict cardiac amyloidosis. *International Journal of Environmental Research and Public Health*, *18*(3), p.908.

[9] Porto, B.M., 2024. Improving triage performance in emergency departments using machine learning and natural language processing: a systematic review. *BMC Emergency Medicine*, *24*(1), p.219.

[10] Ali, H., 2022. Reinforcement learning in healthcare: optimizing treatment strategies, dynamic resource allocation, and adaptive clinical decision-making. *Int J Comput Appl Technol Res*, *11*(3), pp.88-104.

[11] Golovenkin, S.E., Bac, J., Chervov, A., Mirkes, E.M., Orlova, Y.V., Barillot, E., Gorban, A.N. and Zinovyev, A., 2020. Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data. *GigaScience*, *9*(11), p.giaa128.

[12] Subathra, R. and Sumathy, V., 2025. A smart CardioSenseNet framework with advanced data processing models for precise heart disease detection. *Computers in Biology and Medicine*, *185*, p.109473.

[13] Adlung, L., Cohen, Y., Mor, U. and Elinav, E., 2021. Machine learning in clinical decision making. *Med*, *2*(6), pp.642-665.

[14] Desai, R.J., Wang, S.V., Vaduganathan, M., Evers, T. and Schneeweiss, S., 2020. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records

to predict heart failure outcomes. *JAMA network open*, *3*(1), pp.e1918962-e1918962.

[15] UCI Machine Learning Repository. *Myocardial Infarction Complications*.https://archive.ics.uci.edu/dataset/579/myocardial+infarction+complications.

[16] Golovenkin, S.E., Bac, J., Chervov, A., Mirkes, E.M., Orlova, Y.V., Barillot, E., Gorban, A.N. and Zinovyev, A., 2020. Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data. *GigaScience*, *9*(11), p.giaa128.

[17] Amanda, R. and Negara, E.S., 2020. Analysis and implementation machine learning for youtube data classification by comparing the performance of classification algorithms. *Jurnal Online Informatika*, *5*(1), pp.61-72.

[18] Nabi, Z., 2016. Machine learning at scale. In *Pro Spark Streaming: The Zen of Real-Time Analytics Using Apache Spark* (pp. 177-198). Berkeley, CA: Apress.

[19] Jia, B.B. and Zhang, M.L., 2021, July. Multi-dimensional classification via sparse label encoding. In *International Conference on Machine Learning* (pp. 4917-4926). PMLR.

[20] Rattan, V., Mittal, R., Singh, J. and Malik, V., 2021, March. Analyzing the application of SMOTE on machine learning classifiers. In *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)* (pp. 692-695). IEEE.

[21] Liberty, E., Lang, K. and Shmakov, K., 2016, June. Stratified sampling meets machine learning. In *International conference on machine learning* (pp. 2320-2329). PMLR.

[22] Muthukrishnan, R. and Rohini, R., 2016, October. LASSO: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE international conference on advances in computer applications (ICACA)* (pp. 18-20). Ieee.

[23] Abbas, F., Cai, Z., Shoaib, M., Iqbal, J., Ismail, M., Ullah, A.R.I.F., Alrefaei, A.F. and Albeshr, M.F., 2024. *Uncertainty Analysis of Predictive Models for Water Quality Index: Comparative Analysis of XGBoost, Random Forest, SVM, KNN, Gradient Boosting, and Decision Tree Algorithms* [online].

[24] Gönen, M., 2006. Receiver operating characteristic (ROC) curves. *SAS Users Group International (SUGI)*, *31*, pp.210-231.

[25] De Hond, A.A., Steyerberg, E.W. and Van Calster, B., 2022. InterpAreting area under the receiver operating characteristic curve. *The Lancet Digital Health*, *4*(12), pp.e853-e855.

[26] Zhao, S., Yang, Z., Musa, S.S., Ran, J., Chong, M.K., Javanbakht, M., He, D. and Wang, M.H., 2021. Attach importance of the bootstrap t test against Student's t test in clinical epidemiology: a demonstrative comparison using COVID-19 as an example. *Epidemiology & Infection*, *149*, p.e107.

[27] Fan, J., Shi, S., Xiang, H., Fu, L., Duan, Y., Cao, D. and Lu, H., 2024. Predicting Elimination of Small-Molecule Drug Half-Life in Pharmacokinetics Using Ensemble and Consensus Machine Learning Methods. *Journal of Chemical Information and Modeling*, *64*(8), pp.3080-3092.

[28] Beauxis-Aussalet, E. and Hardman, L., 2014, November. Simplifying the visualization of confusion matrix. In *26th Benelux conference on artificial intelligence (BNAIC)*.